Adaptive estimation of the conditional intensity of marker-dependent counting processes

Fabienne Comte, Stéphane Gaïffas, Agathe Guilloux

ADAPTIVE ESTIMATION OF THE CONDITIONAL INTENSITY OF MARKER-DEPENDENT COUNTING PROCESSES

F. COMTE[(1)], S. GAÏFFAS[(2)] & A. GUILLOUX[(3)]

Abstract. We propose in this work an original estimator of the conditional intensity of a marker-dependent counting process, that is, a counting process with covariates. We use model selection methods and provide a non asymptotic bound for the risk of our estimator on a compact set. We show that our estimator reaches automatically a convergence rate over a functional class with a given (unknown) anisotropic regularity. Then, we prove a lower bound which establishes that this rate is optimal. Lastly, we provide a short illustration of the way the estimator works in the context of conditional hazard estimation.

July 12, 2010

AMS (2000) subject classification. 62N02, 62G05.

Keywords. Marker-dependent counting process. Conditional intensity. Model selection. Adaptive estimation. Minimax and Nonparametric methods. Censored data. Conditional hazard function.

1. Introduction

As counting processes can model a great diversity of observations, especially in medicine, actuarial science or economics, their statistical inference has received a continuous attention since half a century - see Andersen et al. (1993) for the most detailed presentation on the subject. In this paper, we propose a new strategy, based on model selection, for the inference for counting processes in presence of covariates. The model considered can be described as follows.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ a filtration satisfying the usual conditions. Let N be a marker-dependent counting process, with compensator Λ with respect to $(\mathcal{F}_t)_{t\geq 0}$, such that $N - \Lambda = M$, where M is a $(\mathcal{F}_t)_{t\geq 0}$-martingale. We assume that N is a marker-dependent counting process satisfying the Aalen multiplicative intensity model in the sense that :

$$\Lambda(t) = \int_0^t \alpha(X, z)Y(z)dz, \text{ for all } t \geq 0 \tag{1}$$

where X is a vector of covariates in $\mathbb{R}^d$ which is $\mathcal{F}_0$-measurable, the process Y is nonnegative and predictable and α is an unknown deterministic function called intensity.

The purpose of this paper is to estimate the intensity function α on the basis of the observation of a n-sample $(X_i, N^i(z), Y^i(z), z \leq \tau)$ for $i = 1, \ldots, n$, where $\tau < +\infty$.

(1) MAP5, University Paris Descartes, France. email: fabienne.comte@parisdescartes.fr,
(2) LSTA University Pierre et Marie Curie, France. email: stephane.gaiffas@upmc.fr,
(3) LSTA University Pierre et Marie Curie, France. email: agathe.guilloux@upmc.fr.

There are many examples, crucial in practice, which fulfill this model. For the seek of conciseness, we restrict our presentation to the three following ones.

Example 1 (Regression model for right-censored data). Let T be a nonnegative random variable (r.v.) with cumulative distribution functions (c.d.f.) F_T, and X a vector of covariates in $\mathbb{R}^d$. We consider in addition that T can be censored. We introduce the nonnegative r.v. C, with c.d.f. G, such that the observable r.v. are $Z = T \wedge C$, $\delta = \mathbb{1}(T \leq C)$ and X. We assume that:

$$(\mathcal{C}) : \ T \text{ and } C \text{ are independent conditionally to } X.$$

In this case, the processes to consider (see e.g. Andersen et al. (1993)) are given, for $i = 1, \dots, n$ and $z \geq 0$, by:

$$N^i(z) = \mathbb{1}(Z_i \leq z, \delta_i = 1) \text{ and } Y^i(z) = \mathbb{1}(Z_i \geq z).$$

The unknown intensity function α to be estimated is the conditional hazard rate of the r.v. T given $X = x$ defined, for all $z > 0$ by:

$$\alpha(x, z) = \alpha_{T|X}(x, z) = \frac{f_{T|X}(x, z)}{1 - F_{T|X}(x, z)},$$

where $f_{T|X}$ and $F_{T|X}$ are respectively the conditional probability density function (p.d.f.) and the conditional c.d.f. of T given X.

Nonparametric estimation of the hazard rate in presence of covariates was initiated by Beran (1981). Stute (1986), Dabrowska (1987), McKeague and Utikal (1990) and Li and Doss (1995) extended his results. Many authors have considered semiparametric estimation of the hazard rate, beginning with Cox (1972), see Andersen et al. (1993) for a review of the enormous literature on semiparametric models. We refer to Huang (1999) and Linton et al. (2003) for some recent developments.

Adaptive nonparametric estimation for censored data in presence of covariates has been considered by LeBlanc and Crowley (1999) or Castellan and Letué (2000) for particular functional Cox models: in these works, $\alpha(x, z) = \exp(f(x))\alpha_0(z)$, only f is estimated. On the other hand, Brunel et al. (2007) constructed an optimal adaptive estimator of the conditional density in a general model.

Example 2 (Cox processes). Let η^i, for $i = 1, \dots, n$, be a Cox process (see Kaar (1986)) on $\mathbb{R}_+$ with random mean-measure Λ^i given by :

$$\Lambda^i(t) = \int_0^t \alpha(X_i, z) dz,$$

where X_i is a vector of covariates in $\mathbb{R}^d$. In this context the predictable process Y of Equation (1) constantly equals 1. As a consequence, these processes can be seen as generalizations of nonhomogeneous Poisson processes on $\mathbb{R}_+$ with random intensities. This is a particular case of longitudinal data, see e.g. Example VII.2.15 in Andersen et al. (1993). The nonparametric estimation of the intensity of Poisson processes without covariates has been considered in several papers. We refer to Reynaud-Bouret (2003) and Baraud and Birgé (2009) for the adaptive estimation of the intensity of nonhomogeneous Poisson processes in general spaces.

Example 3 (Regression model for transition intensities of Markov processes). Consider a n-sample of nonhomogeneous time-continuous Markov processes $P^1, \dots, P^n$ with finite state space $\{1, \dots, k\}$ and denote by α_{jl} the transition intensity from state j to state l. For individual i with covariate X_i, let $N^i_{jl}(t)$ be the number of observed direct transitions from j to l before time t (we allow the possibility of right-censoring for example). Conditionally on the initial state, the counting process N^i_{jl} verifies the following Aalen multiplicative intensity model:

$$N^i_{jl}(t) = \int_0^t \alpha_{jl}(X_i, z) Y^i_j(z) dz + M^i(t) \text{ for all } t \geq 0,$$

where $Y^i_j(t) = \mathbb{1}\{P^i(t-) = j\}$ for all $t \geq 0$, see Andersen et al. (1993) or Jacobsen (1982). This setting is discussed in Andersen et al. (1993), see Example VII.11 on mortality and nephropathy for insulin dependent diabetics.

We finally cite three papers, where different strategies for the estimation of the intensity of counting processes is considered, gathering as a consequence all the previous examples, but in none of them the presence of covariates was considered. Ramlau-Hansen (1983) proposed a kernel-type estimator, Grégoire (1993) studied cross-validation for these estimators. More recently, Reynaud-Bouret (2006) considered adaptive estimation by model selection.

Our aim in this work is to provide an optimal adaptive nonparametric estimator of the conditional intensity. Our estimation procedure involves the minimization of a so-called contrast. To achieve that purpose, we proceed as follows. In Section 2, we describe the estimation procedure: we explain how the contrast is built, on which collections of spaces the estimators are defined and how the relevant space is selected via a data driven penalized criterion. In Section 3, we state oracle inequalities for our estimator (see Theorems 1 and 2), a resulting upper bound (see Corollary 1) and a lower bound (see Theorem 3), the latter asserts the optimality in the minimax sense. The examples of Section 4 are taken in the setting of Example 1, in order to provide a short illustration of the practical properties of our estimator. Lastly, proofs are gathered in Sections 5 and 6.

Remark 1. An inherent remark about this model is that there is no reason for the conditional intensity $\alpha(x, z)$ to have the same behavior with respect to the z (time) and x (covariates) variables. This is the reason why it is mandatory in our purely nonparametric setting to consider anisotropic regularity for α. Think for instance of the very popular case of proportional hazards Cox model, see Cox (1972), it is assumed that $\alpha(x, z) = \alpha_0(z) \exp(\beta^\top x)$ for some unknown function α_0 and unknown vector $\beta \in \mathbb{R}^d$. Of course, in this model, the smoothness in the x direction is higher than in the z direction.

For the sake of simplicity, we will assume in the following that the covariate X is one-dimensional.

2. Description of the procedure

Our estimation procedure involves the minimization of a contrast. This contrast is tuned to the problem considered in this paper, as explained in the next section.

2.1. **Definition of the contrast.** Let $A = A_1 \times [0,\tau]$ be a compact set of $\mathbb{R} \times \mathbb{R}_+$ on which the function α will be estimated. Without loss of generality, we set $A = [0,1]\times[0,\tau]$. Let h be a function in $(L^2 \cap L^\infty)(A)$. Define the contrast function:

$$\gamma_n(h) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau h^2(X_i,z)Y^i(z)dz - \frac{2}{n}\sum_{i=1}^n \int_0^\tau h(X_i,z)dN^i(z). \tag{2}$$

This contrast is of least-squares type adapted to the problem considered here. Since each N^i admits a Doob-Meyer decomposition ($N^i = \Lambda^i + M^i$), we have:

$$\gamma_n(h) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau h^2(X_i,z)Y^i(z)dz - \frac{2}{n}\sum_{i=1}^n \int_0^\tau h(X_i,z)d\Lambda^i(z) - \frac{2}{n}\sum_{i=1}^n \int_0^\tau h(X_i,z)dM^i(z),$$

so that:

$$\mathbb{E}\big(\gamma_n(h)\big) = \mathbb{E}\big(\int_0^\tau h^2(X,z)Y(z)dz\big) - \mathbb{E}\big(2\int_0^\tau h(X,z)d\Lambda(z)\big).$$

Let F_X denote the c.d.f. of the covariate X and $\|\cdot\|_\mu$ the norm defined by:

$$\|h\|_\mu^2 \;:=\; \mathbb{E}\big(\int_0^\tau h^2(X,z)Y(z)dz\big) = \iint_A h^2(x,z)d\mu(x,z),$$

where $d\mu(x,z) := \mathbb{E}(Y(z)|X=x)F_X(dx)dz$. By the Aalen multiplicative intensity model, see Equation (1), we get:

$$\mathbb{E}\big(\gamma_n(h)\big) = \|h\|_\mu^2 - 2\iint h(x,z)\alpha(x,z)\mathbb{E}(Y(z)|X=x)F_X(dx)dz = \|h-\alpha\|_\mu^2 - \|\alpha\|_\mu^2.$$

This explains why minimizing $\gamma_n(\cdot)$ over an appropriate set of functions described below, is a relevant strategy to estimate α.

Example 1 continued. In the particular case of regression for right-censored data, the conditional hazard function is estimated and the contrast function has the following form:

$$\gamma_n(h) \;=\; \frac{1}{n}\sum_{i=1}^n \int_0^\tau h^2(X_i,z)\mathbb{1}(Z_i \geq z)dz - \frac{2}{n}\sum_{i=1}^n \delta_i h(X_i,Z_i).$$

We have in addition an explicit formula for $d\mu(x,z)$:

$$d\mu(x,z) = (1 - L_{Z|X}(z,x))F_X(dx)dz, \tag{3}$$

where

$$1 - L_{Z|X}(z,x) := \mathbb{P}(Z \geq z|X=x) = (1-F_{T|X}(x,z))(1-G_{C|X}(x,z))$$

and $G_{C|X}$ is the conditional c.d.f. of C given X.

Remark 2. In our setting, it is possible to let the censoring depend on the covariates, as in Dabrowska (1989) or, more recently Heuchenne and Van Keilegom (2006). Assumption $(\mathcal{C})$ above is weaker than the assumption: T and C are independent and $\mathbb{P}(T \leq C|X,T) = \mathbb{P}(T \leq C|T)$ in Stute (1996). See Delecroix et al. (2008), p.249, for further discussions on this matter.

2.2. **Assumptions and notations.** Before defining the estimation procedure, we need to introduce some assumptions and notations. Define the norms

$$\|h\|_A^2 := \iint_A h^2(x,z)dxdz \text{ and } \|h\|_{\infty,A} := \sup_{(x,z)\in A} |h(x,z)|,$$

and assume that the following condition holds:

- $(\mathcal{A}1)$ The covariates X_i admit a p.d.f. f_X such that $\sup_{A_1} |f_X| \leq f_1 < +\infty$.

Assumption $(\mathcal{A}1)$ implies that μ admits a density w.r.t. the Lebesgue measure. We denote by f this density:

$$d\mu(x,z) = f(x,z)dxdz \text{ where } f(x,z) = \mathbb{E}(Y(z)|X=x)f_X(x). \tag{4}$$

We also assume:

- $(\mathcal{A}2)$ There exists $f_0 > 0$, such that $\forall (x,z) \in A_1 \times [0,\tau],\ f(x,z) \geq f_0$.
- $(\mathcal{A}3)$ $\forall (x,z) \in A_1 \times [0,\tau],\ \alpha(x,z) \leq \|\alpha\|_{\infty,A} < +\infty$.
- $(\mathcal{A}4)$ $\forall i, \forall t,\ Y^i(t) \leq C_Y$ where C_Y is a known fixed constant.

Remark 3. Assumption $(\mathcal{A}2)$ is fulfilled if Y is bounded from below in expectation and if f_X is bounded from below. The requirement that the density of the design is bounded away from zero is standard in a regression model, for instance. Assumption $(\mathcal{A}2)$ reduces to such a condition in Example 2 (Cox processes), where we have $f(z,x) = I(z \in [0,\tau])f_X(x)$. In the general setting of counting processes, a lower bound on the expectation of Y is classical, see Reynaud-Bouret (2006) p.648. In the censored case (Example 1), we can write:

$$\mathbb{E}(Y(z)|X=x) = \mathbb{E}(\mathbb{1}(T \wedge C \geq z)|X=x) = (1 - F_{T|X}(x,z))(1 - G_{C|X}(x,z-)).$$

It is a well-known fact (see e.g. Andersen et al. (1993), p 193-194) that the Kaplan-Meier estimator is consistent, for each x (with no further assumption) only on intervals of the form $[0,\tau_x]$, where $\tau_x < \sup\{s \geq 0, (1 - F_{T|X}(x,s))(1 - G_{C|X}(x,s)) > 0\}$. We can take $\tau = \inf_{x\in[0,1]} \tau_x$. In view of (3), this justifies our Assumption $(\mathcal{A}2)$ in this case.

Lastly, in the examples described in Section 1, Assumption $(\mathcal{A}4)$ is clearly fulfilled with $C_Y = 1$. We will set $C_Y = 1$ in the following for simplicity. This implies together with $(\mathcal{A}1)$ that $\forall (x,z) \in A, |f(x,z)| \leq f_1$.

2.3. **Definition of the estimator.** We use the usual model selection paradigm (see, for instance, Massart (2007)): first minimize the contrast $\gamma_n(\cdot)$ over a finite-dimensional function space S_m, then select the appropriate space by penalization. We introduce a collection $\{S_m : m \in \mathcal{M}_n\}$ of projection spaces: S_m is called a model and $\mathcal{M}_n$ is a set of multi-indexes (see the examples in Section 2.4). For each $m = (m_1, m_2)$, the space S_m of functions with support in $A = [0,1] \times [0,\tau]$ is defined by:

$$S_m = F_{m_1} \otimes H_{m_2} = \Big\{ h : h(x,z) = \sum_{j\in J_m} \sum_{k\in K_m} a_{j,k}^m \varphi_j^m(x)\psi_k^m(z),\ a_{j,k}^m \in \mathbb{R} \Big\},$$

where F_{m_1} and H_{m_2} are subspaces of $(L^2 \cap L^\infty)(A_1)$ and $(L^2 \cap L^\infty)([0,\tau])$ respectively spanned by two orthonormal bases $(\varphi_j^m)_{j\in J_m}$ with $|J_m| = D_{m_1}$ and $(\psi_k^m)_{k\in K_m}$ with $|K_m| = D_{m_2}$. For all j and all k, the supports of φ_j^m and ψ_k^m are respectively included in A_1 and $[0,\tau]$. Here j and k are not necessarily integers, they can be pairs of integers, as in the piecewise polynomial or the wavelet cases, see Section 2.4.

Remark 4. From a theoretical point of view, we could consider that the covariates X are in $\mathbb{R}^d$. For this end, we would have to consider models of the form $S_m = F_{m_1} \otimes \cdots \otimes F_{m_d} \otimes H_{m_{d+1}}$. However, this would make the proofs more intricate. Note also that the convergence rate would be slower because of the curse of dimensionality. For the sake of clarity, we restrict ourselves to $X \in \mathbb{R}$.

The first step would be to define $\hat{\alpha}_m = \operatorname{argmin}_{h \in S_m} \gamma_n(h)$. To that end, let $h(x,y) = \sum_{j \in J_m} \sum_{k \in K_m} a_{j,k} \varphi_j^m(x) \psi_k^m(y)$ be a function in S_m. To compute $\hat{\alpha}_m$, we have to solve:

$$\forall j_0 \forall k_0, \quad \frac{\partial \gamma_n(h)}{\partial a_{j_0,k_0}} = 0 \Leftrightarrow G_m A_m = \Upsilon_m,$$

where A_m denotes the matrix $(a_{j,k})_{j \in J_m, k \in K_m}$,

$$G_m := \Big(\frac{1}{n} \sum_{i=1}^n \varphi_j^m(X_i) \varphi_l^m(X_i) \int_0^\tau \psi_k^m(z) \psi_p^m(z) Y^i(z) dz \Big)_{(j,k),(l,p) \in J_m \times K_m}$$

and

$$\Upsilon_m := \Big(\frac{1}{n} \sum_{i=1}^n \varphi_j^m(X_i) \int_0^\tau \psi_k^m(z) dN^i(z) \Big)_{j \in J_m, k \in K_m}.$$

Unfortunately G_m may not be invertible. To overcome this problem, we modify the definition of $\hat{\alpha}_m$ in the following way:

$$\hat{\alpha}_m := \begin{cases} \operatorname{argmin}_{h \in S_m} \gamma_n(h) & \text{on } \hat{\Gamma}_m \\ 0 & \text{on } \hat{\Gamma}_m^{\complement} \end{cases}, \tag{5}$$

where

$$\hat{\Gamma}_m := \Big\{ \min \mathrm{Sp}(G_m) \geq \max(\hat{f}_0/3, n^{-1/2}) \Big\}$$

where $\mathrm{Sp}(G_m)$ denotes the spectrum of G_m i.e. the set of the eigenvalues of the matrix G_m (it is easy to see that they are nonnegative). The estimator $\hat{f}_0$ of f_0 (the minimum of the density f, see $(\mathcal{A}2)$) is required to fulfill the following assumption:

- $(\mathcal{A}5)$ For any integer $k \geq 1$, there are positive constants C_0 and n_0 such that

$$\mathbb{P}(|\hat{f}_0 - f_0| > f_0/2) \leq C_0/n^k \quad \text{for any} \quad n \geq n_0.$$

An estimator satisfying $(\mathcal{A}5)$ is defined in Section 3.5, where the constants C_0 and n_0 depend on $k, f_0, f_1, \tau, \phi_1, \phi_2$. In fact, $k = 7$ is enough for the proofs. We refer the reader to the proof of Lemma 1, see Section 6, for an explanation of the presence of $n^{1/2}$ in the definition of $\hat{\Gamma}_m$. In practice, this constraint is generally not used (the matrix is invertible, otherwise another model is considered).

The final step is to select the relevant space via the penalized criterion:

$$\hat{m} = \operatorname*{argmin}_{m \in \mathcal{M}_n} \Big(\gamma_n(\hat{\alpha}_m) + \mathrm{pen}(m) \Big), \tag{6}$$

where $\mathrm{pen}(m)$ is defined in Theorem 1 below, see Section 3. Our estimator of α on A is then $\hat{\alpha}_{\hat{m}}$.

2.4. **Assumptions on the models and examples.** Let us introduce the following set of assumptions on the models $\{S_m : m \in \mathcal{M}_n\}$, which are usual in model selection techniques.

- ($\mathcal{M}1$) For $i = 1, 2$, $\mathcal{D}_n^{(i)} := \max_{m\in\mathcal{M}_n} D_{m_i} \leq n^{1/4}/\sqrt{\log n}$. We shall denote by $\mathcal{F}_n$ (respectively $\mathcal{H}_n$) the space with dimension $\mathcal{D}_n^{(1)}$ (resp. $\mathcal{D}_n^{(2)}$).
- ($\mathcal{M}2$) There exist $\phi_1 > 0, \phi_2 > 0$ such that, for all u in F_{m_1} and for all v in H_{m_2}, we have
$$\sup_{x\in A_1} |u(x)|^2 \leq \phi_1 D_{m_1} \int_{A_1} u^2 \text{ and } \sup_{x\in[0,\tau]} |v(x)|^2 \leq \phi_2 D_{m_2} \int_{[0,\tau]} v^2.$$
By letting $\phi_0 = \sqrt{\phi_1\phi_2}$, that leads to
$$\forall h \in S_m \qquad \|h\|_{\infty,A} \leq \phi_0 \sqrt{D_{m_1} D_{m_2}} \|h\|_A. \tag{7}$$
- ($\mathcal{M}3$) Nesting condition:
$$D_{m_1} \leq D_{m_1'} \Rightarrow F_{m_1} \subset F_{m_1'} \text{ and } D_{m_2} \leq D_{m_2'} \Rightarrow H_{m_2} \subset H_{m_2'}.$$
Moreover, there exists a global nesting space $\mathcal{S}_n = \mathcal{F}_n \otimes \mathcal{H}_n$ in the collection, such that $\forall m \in \mathcal{M}_n, S_m \subset \mathcal{S}_n$ and $\dim(\mathcal{S}_n) := N_n \leq \sqrt{n/\log n}$.

Remark 5. We emphasize that ϕ_2 depends on τ and is in most examples proportional to $1/\tau$.

Assumptions ($\mathcal{M}1$)–($\mathcal{M}3$) are not too restrictive. Indeed, they are verified for the spaces F_{m_1} (and H_{m_2}) on $A_1 = [0, 1]$ spanned by the following bases (see Barron et al. (1999)):

- $[T]$ Trigonometric basis: $\text{span}(\varphi_0, \ldots, \varphi_{m_1-1})$ with $\varphi_0 = \mathbb{1}([0, 1])$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\ \mathbb{1}([0, 1])(x)$, $\varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)\mathbb{1}([0, 1])(x)$ for $j \geq 1$. For this model $D_{m_1} = m_1$ and $\phi_1 = 2$ hold.
- $[DP]$ Regular piecewise polynomial basis: polynomials of degree $0, \ldots, r$ (where r is fixed) on each interval $[(l-1)/2^D, l/2^D[$ with $l = 1, \ldots, 2^D$. In this case, we have $m_1 = (D, r)$, $J_m = \{j = (l, d),\ 1 \leq l \leq 2^D, 0 \leq d \leq r\}$, $D_{m_1} = (r + 1)2^D$ and $\phi_1 = \sqrt{r + 1}$.
- $[W]$ Wavelet basis on an interval: $\text{span}(\Psi_{j,k} : j = l - 1, \ldots, m_1, k \in \Lambda(j))$, where l and m_1 are integers (l corresponds to the number of vanishing moments of the basis). The $\Psi_{j,k}$ are, depending on the localization parameter k, either translations and dilatations of a pair $\{\phi, \psi\}$ of scaling function and wavelet with a compact support, or so-called edge scaling functions and wavelets. We give more details in Appendix A.1. By construction, the elements of this basis have their supports included in A_1, and they have as many vanishing moments as ψ.
- $[H]$ Histogram basis: for $A_1 = [0, 1]$, $\text{span}(\varphi_1, \ldots, \varphi_{2^{m_1}})$ with $\varphi_j = 2^{m_1/2}\mathbb{1}([(j - 1)/2^{m_1}, j/2^{m_1}[)$ for $j = 1, \ldots, 2^{m_1}$. Here $D_{m_1} = 2^{m_1}$, $\phi_1 = 1$. Notice that $[H]$ is a particular case of both $[DP]$ and $[W]$.

Clearly, if $\varphi_1, \ldots, \varphi_D$ is an orthonormal basis in $L^2([0, 1])$, then $\tau^{-1/2}\varphi_1(\cdot/\tau), \ldots, \tau^{-1/2}\varphi_D(\cdot/\tau)$ is an orthonormal basis in $L^2([0, \tau])$.

Remark 6. The first assumption ($\mathcal{M}1$) prevents the dimension from being too large compared to the number of observations. We can relax considerably this constraint for localized basis: for histogram basis, piecewise polynomial basis and wavelets, ($\mathcal{M}1$) can be relaxed to the weaker condition: $\mathcal{D}_n^{(i)} \leq \sqrt{n/\log n}$. Analogously in ($\mathcal{M}3$), we would

get $N_n \leq n/\log n$. The condition $(\mathcal{M}2)$ implies a useful link between the L^2 norm and the infinite norm. The third assumption $(\mathcal{M}3)$ implies in particular that $\forall m, m' \in \mathcal{M}_n$, $S_m + S_{m'} \subset \mathcal{S}_n$. This condition is useful for the chaining argument used in the proofs, see Section 6.4.

3. Main results

3.1. **Oracle inequality.** We define α_m as the orthogonal projection of $\alpha\mathbb{1}(A)$ on S_m. The estimator $\hat{\alpha}_{\hat{m}}$ where $\hat{\alpha}_m$ is given by (5) and $\hat{m}$ is given by (6) satisfies the following oracle inequality.

Theorem 1. *Let* $(\mathcal{A}1)$ – $(\mathcal{A}5)$ *and* $(\mathcal{M}1)$ – $(\mathcal{M}3)$ *hold. Define the following penalty*:

$$\mathrm{pen}(m) := K_0(1 + \|\alpha\|_{\infty,A})\frac{D_{m_1} D_{m_2}}{n}, \tag{8}$$

where K_0 *is a numerical constant. We have*

$$\mathbb{E}(\|\alpha\mathbb{1}(A) - \hat{\alpha}_{\hat{m}}\|_\mu^2) \leq \kappa_0 \inf_{m\in\mathcal{M}_n}\{\|\alpha\mathbb{1}(A) - \alpha_m\|_\mu^2 + \mathrm{pen}(m)\} + \frac{C}{n} \tag{9}$$

for any $n \geq n_0$*, where* n_0 *is a constant coming from Assumption* $(\mathcal{A}5)$ *(see Section 2.3), where* κ_0 *is a numerical constant and* C *is a constant depending on* $\phi_1, \phi_2, \|\alpha\|_{\infty,A}, f_0, f_1$ *and* τ*.*

The proof of Theorem 1 involves a deviation inequality for the empirical process

$$\nu_n(h) := \frac{1}{n}\sum_{i=1}^n \int_0^\tau h(X_i, z) dM^i(z),$$

where $M^i(t) = N^i(t) - \int_0^t \alpha(X_i, z)Y^i(z)dz$ are martingales, see Section 1, and a $L^2 - L^\infty$ chaining argument.

3.2. **Adaptive upper bound.** From Theorem 1, we can derive the rate of convergence of $\hat{\alpha}_{\hat{m}}$ over anisotropic Besov spaces. We recall that anisotropy is almost mandatory in this context, see Remark 1. For that purpose, assume that α restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ on A with regularity $\boldsymbol{\beta} = (\beta_1, \beta_2)$. Let us recall the definition of $B^{\boldsymbol{\beta}}_{2,\infty}(A)$. Let $\{e_1, e_2\}$ the canonical basis of $\mathbb{R}^2$ and take $A^r_{h,i} := \{x \in \mathbb{R}^2; x, x + he_i, \ldots, x + rhe_i \in A\}$, for $i = 1, 2$. For $x \in A^r_{h,i}$, let

$$\Delta^r_{h,i} g(x) = \sum_{k=0}^r (-1)^{r-k}\binom{r}{k} g(x + khe_i)$$

be the rth difference operator with step h. For $t > 0$, the directional moduli of smoothness are given by

$$\omega_{r_i,i}(g, t) = \sup_{|h|\leq t}\Big(\int_{A^{r_i}_{h,i}} |\Delta^{r_i}_{h,i} g(x)|^2 dx\Big)^{1/2}.$$

Consider the Besov norm

$$\|\alpha\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} := \|\alpha\|_A + |\alpha|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} = \|\alpha\|_A + \sup_{t>0}\sum_{i=1}^2 t^{-\beta_i}\omega_{r_i,i}(g, t), \tag{10}$$

and define the Besov space $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ as the set of functions g such that $\|g\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} < +\infty$, and for $L > 0$, consider the ball

$$B^{\boldsymbol{\beta}}_{2,\infty}(A, L) = \{\alpha \in B^{\boldsymbol{\beta}}_{2,\infty}(A) : \|\alpha\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} \leq L\}.$$

More details concerning Besov spaces can be found in Triebel (2006). The next corollary shows that $\hat{\alpha}_{\hat{m}}$ adapts to the unknown anisotropic smoothness of α.

Corollary 1. *Assume that α restricted to A belongs to $B^{\boldsymbol{\beta}}_{2,\infty}(A, L)$, with smoothness $\boldsymbol{\beta} = (\beta_1, \beta_2)$ such that $\beta_1 > 1/2$ and $\beta_2 > 1/2$. We consider the piecewise polynomial or wavelet spaces described in Subsection 2.4 (with the regularity of the polynomials and the wavelets larger than $\beta_i - 1$). Then, under the assumptions of Theorem 1, we have*

$$\mathbb{E}\|\alpha - \hat{\alpha}_{\hat{m}}\|^2_A \leq C n^{-\frac{2\bar{\beta}}{2\bar{\beta}+2}}$$

where $\bar{\beta}$ is the harmonic mean of β_1 and β_2 (i.e. $2/\bar{\beta} = 1/\beta_1 + 1/\beta_2$) and C depends on L, τ, ϕ_0, f_0, f_1 and $\|\alpha\|_{\infty,A}$.

The rate of convergence achieved by $\hat{\alpha}_{\hat{m}}$ in Corollary 1 is optimal in the minimax sense as proved in Theorem 3 below. For trigonometric spaces, the result also holds, but for $\beta_1 > 3/2$ and $\beta_2 > 3/2$ (because of $(\mathcal{M}1)$).

Moreover, assuming for example that $\beta_2 > \beta_1$, one can see in the proof of Corollary 1 that the estimator chooses a space of dimension $D_{\hat{m}_2} = D^{\beta_1/\beta_2}_{\hat{m}_1} < D_{\hat{m}_1}$. This shows that the estimator is adaptive with respect to the approximation space for each directional regularity.

3.3. **Random penalty.** It is worth noting that the penalty defined in Equation (8) involves the unknown quantity $\|\alpha\|_{\infty,A}$. This problem occurs occasionally in penalization procedures, see for instance Comte (2001) or Lacour (2007a). The solution is to replace it by an estimator:

$$\widehat{\text{pen}}(m) = K_1(1 + \|\hat{\alpha}_{m*}\|_{A,\infty})\frac{D_{m_1} D_{m_2}}{n}, \tag{11}$$

where K_1 is a numerical constant and $\hat{\alpha}_{m^*}$ is a rough estimator of α computed on an arbitrary space S_{m^*} with dimension $D_{m^*} = D_{m^*_1} D_{m^*_2}$. Let us consider

$$\hat{\hat{m}} = \arg\min_{m \in \mathcal{M}_n} \left(\gamma_n(\hat{\alpha}_m) + \widehat{\text{pen}}(m)\right). \tag{12}$$

Then we can prove the following result:

Theorem 2. *Let the assumptions of Theorem 1 be satisfied. Consider the estimator $\hat{\alpha}_{\hat{\hat{m}}}$ defined by (5)-(12)-(11), where the term $\hat{\alpha}_{m^*}$ is computed with (5) on a space S_{m^*} in collection [T] with dimension D_{m^*} such that*

$$D_{m^*_1} = D_{m^*_2} = n^{1/4}.$$

If α restricted to A belongs to the anisotropic Besov space $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ with regularity $\boldsymbol{\beta} = (\beta_1, \beta_2)$ such that $\beta_1 > 2$ and $\beta_2 > 2$, then, for n large enough,

$$\mathbb{E}(\|\alpha \mathbb{1}(A) - \hat{\alpha}_{\hat{\hat{m}}}\|^2_\mu) \leq \kappa_1 \inf_{m \in \mathcal{M}_n} \{\|\alpha \mathbb{1}(A) - \alpha_m\|^2_\mu + (1 + \|\alpha\|_{\infty,A})\frac{D_{m_1} D_{m_2}}{n}\} + \frac{C}{n} \tag{13}$$

where κ_1 is a numerical constant and C is a constant depending on $\phi_1, \phi_2, \|\alpha\|_{\infty,A}, f_0, f_1$ and τ.

Obviously, we can deduce from Theorem 2 a Corollary similar to Corollary 1 concerning the asymptotic rate of the estimator on Besov balls.

3.4. **Lower bound.** In the next Theorem, we prove that the rate $n^{-2\bar{\beta}/(2\bar{\beta}+2)}$ is optimal over $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ where we recall that $2/\bar{\beta} = 1/\beta_1 + 1/\beta_2$. Recall that the Besov ball $B^{\boldsymbol{\beta}}_{2,\infty}(A, L)$ is defined in Section 3.2. Let us denote by $\mathbb{E}_\alpha$ the integration w.r.t. the joint law $\mathbb{P}_\alpha$, when the intensity is α, of the n-sample $(X_i, N^i(z), Y^i(z); z \le \tau, i = 1, \dots, n)$.

Theorem 3. *Assume that Assumption $(\mathcal{A}1)$ holds. Then there is a constant $C > 0$ that depends on $\boldsymbol{\beta}, L, \tau$ and f_1 such that*

$$\inf_{\tilde{\alpha}} \sup_{\alpha \in B^{\boldsymbol{\beta}}_{2,\infty}(A,L)} \mathbb{E}_\alpha \|\tilde{\alpha} - \alpha\|_A^2 \ge C n^{-2\bar{\beta}/(2\bar{\beta}+2)}$$

for n large enough, where the infimum is taken among all estimators.

Remark 7. There is a slight difference between the statements of Theorem 3 and Corollary 1: the upper bound in Corollary 1 needs Assumption $(\mathcal{A}2)$ [which requires that $f(x,z) = \mathbb{E}(Y(z)|X = x)f_X(x) \ge f_0$] while Theorem 3 does not. However Corollary 1 and Theorem 3 are stated on the same functional sets. This kind of difference between the statements of upper and lower bounds is classical, and can be found in regression models as well, see the dicussion in Stone (1980) p.1351 for a regression model.

3.5. **Estimation of f_0.** We recall that f is the density of μ, which is defined in Equation (4). We define

$$\hat{f}_m = \operatorname*{argmin}_{h \in S_m} \upsilon_n(h) \text{ where } \upsilon_n(h) = \|h\|^2 - \frac{2}{n}\sum_{i=1}^n \int_0^\tau h(X_i, z) Y^i(z) dz. \tag{14}$$

This estimator admits a simple explicit formulation:

$$\hat{f}_m(x,z) = \sum_{(j,k) \in J_m \times K_m} \hat{b}_{j,k} \varphi_j^m(x) \psi_k^m(y), \text{ with } \hat{b}_{j,k} = \frac{1}{n}\sum_{i=1}^n \varphi_j^m(X_i) \int_0^\tau \psi_k^m(z) Y^i(z) dz. \tag{15}$$

As before, we consider estimation of f over the compact set $A = [0,1] \times [0,\tau]$. We choose the space H_{m_2} as the space with maximal dimension, as explained below. Let us denote it by $\mathcal{H}_n$, by $\mathcal{D}_n^{(2)} = \dim(\mathcal{H}_n)$ its dimension (see $(\mathcal{M}1)$) and by ℓ_n its index so that $H_{\ell_n} = \mathcal{H}_n$. Hence, we consider, instead of a general $\hat{f}_m$, the estimator

$$\hat{f}_{m_1} := \operatorname*{argmin}_{h \in F_{m_1} \times \mathcal{H}_n} \upsilon_n(h).$$

We are now in a position to define an estimator of f_0 by considering any $\inf_{(x,z)\in A} \hat{f}_{m_1}(x,z)$ with a given m_1. Indeed, an arbitrary choice is sufficient for our estimation problem concerning f_0. In our setting, only a rough estimation of the lower bound on f is useful. Therefore, the estimator $\hat{f}_0$ used in (5) for the construction of $\hat{\alpha}_m$ can be defined by:

$$\hat{f}_0 := \inf_{(x,z)\in A} \hat{f}_{m_1^*}(x,z) \text{ with } D_{m_1^*} = \dim(F_{m_1^*}). \tag{16}$$

Then, the following result holds:

Proposition 1. *Consider* $\hat{f}_0$ *defined by* (16) *in the basis* [T] *with* $D_{m_1^*} = \mathcal{D}_n^{(2)} = n^{1/4}/\sqrt{\log n}$. *Assume that* $f \in \mathcal{B}_{2,\infty}^{(\tilde{\beta}_1,\tilde{\beta}_2)}(A)$ *with* $\tilde{\beta}_1 > 2$, $\tilde{\beta}_2 > 2$. *Then, for any* $k \in \mathbb{N}$, *there are positive constants* n_0 *and* C_0 *such that*

$$\mathbb{P}(|\hat{f}_0 - f_0| > f_0/2) \leq C_0/n^k$$

for any $n \geq n_0$, *where* C_0 *and* n_0 *are constant depending on* k, τ, f_0, f_1, ϕ_1 *and* ϕ_2. *This proves that* $\hat{f}_0$ *fulfills Assumption* $(\mathcal{A}5)$.

The proof of this result is given in Section 6.

4. Illustration

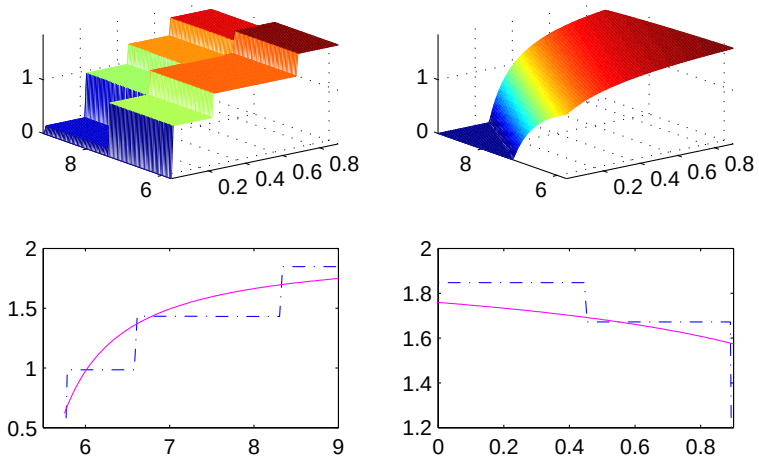

FIGURE 1. Case (NL) Estimated (top left) and true (top right) conditional hazard rates and example of sections (bottom) for a fixed value of x (left) or y (right).

In this section, we give a numerical illustration of the adaptive estimator $\hat{\alpha}_{\hat{m}}$, defined in Section 2, computed with the dyadic histogram basis $[H]$. We sample i.i.d. data $(X_1, T_1), \ldots, (X_n, T_n)$ in three particular cases of the regression model of Example 1 from Section 1. For the sake of simplicity, we simulate the covariates X_i with the uniform distribution on $[0, 1]$. The size of the data set is $n = 1000$.

- Case (NL). Non-Linear regression:

  $$T_i = b(X_i) + \sigma\varepsilon_i.$$

  We simulate ε_i with a $\chi^2(4)$ distribution, $\sigma = 1/4$ and $b(x) = 2x + 5$. Note that in this case, the hazard function to be estimated is $=$

  $$\alpha_{\mathrm{NL}}(x, t) = \frac{1}{\sigma}\alpha_\varepsilon\Big(\frac{t - b(x)}{\sigma}\Big),$$

  where α_ε denotes the hazard function of ε.

- Case (AFT). Accelerated Failure Time model:

$$\log(T_i) = a + bX_i + \varepsilon_i,$$

where the ε_i are standard normal and $a = 5$ and $b = 2$. The hazard function to be estimated is then:

$$\alpha_{AFT}(x,t) = \frac{\alpha_\varepsilon(\log(t) - (a+bx))}{t}.$$

- Case (PH). Proportional Hazards model (see Castellan and Letué (2000), LeBlanc and Crowley (1999)): in this case, the hazard writes

$$\alpha(x,t) = \exp(b(x))\alpha_0(t).$$

We take $b(x) = bx$ with $b = 0.4$ and $\alpha_0(t) = a\lambda t^{a-1}$, which is a Weibull hazard function with $a = 3$ and $\lambda = 1$.

We choose to compute and plot our estimators with histogram bases for two reasons: first, it makes the estimator much easier to compute; secondly, it shows very well how the changes are captured, and when an anisotropic choice is performed by the estimation procedure. More sophisticated implementation is beyond the scope of the paper.

The penalty is taken as

$$\widehat{\mathrm{pen}}(m_1,m_2) = \kappa(1 + \|\hat{\alpha}\|_{\infty,A})\frac{2^{m_1+m_2}}{n},$$

with $\kappa = 4$. Note that, for sake of simplicity, $\|\hat{\alpha}\|_{\infty,A}$ is estimated by $\max_{j,k} \hat{a}_{j,k}$ (the largest histogram coefficients) instead of the trigonometric basis, which was used for technical reasons in Theorem 2: this is because it makes the procedure faster, since all $\hat{a}_{j,k}$ are already computed for estimation. These coefficients are computed on the largest space which is considered (taken with dimension $\sqrt{n}$).

We can see from Figures 1-3 that the algorithm exploits the opportunity (Figures 1 and 3) of choosing different dimensions in the two directions, and that it gives a good account of the general form of the surfaces.

5. Proofs of the main results

5.1. **Proof of Theorem 1.** We define, for h_1, h_2 in $L^2 \cap L^\infty(A)$, the empirical scalar product

$$\langle h_1, h_2\rangle_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau h_1(X_i,z)h_2(X_i,z)Y^i(z)dz\mathbb{1}(X_i \in [0,1]) \tag{17}$$

and the associated empirical norm $\|h_1\|_n^2 = \langle h_1, h_1\rangle_n$ which is such that

$$\mathbb{E}(\|h_1\|_n^2) = \iint_A h_1^2(x,y)d\mu(x,y) = \iint_A h_1^2(x,y)f(x,y)dxdy = \|h_1\|_\mu^2$$

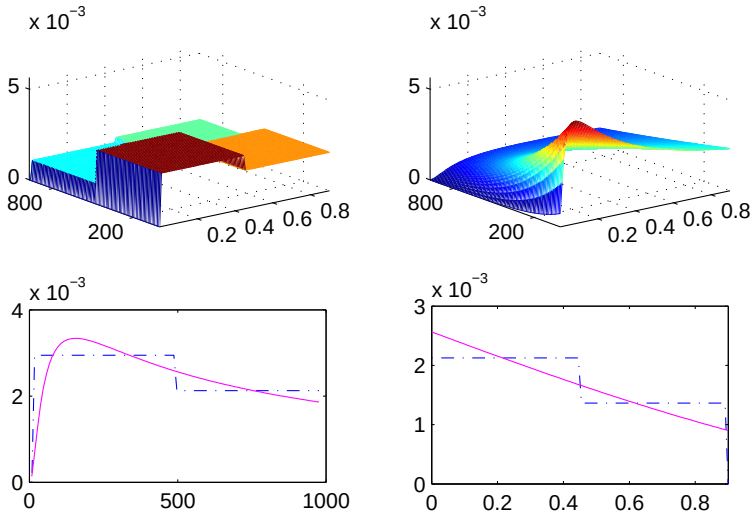

FIGURE 2. Case (AFT) Estimated (top left) and true (top right) conditional hazard rates and example of sections (bottom) for a fixed value of x (left) or y (right).

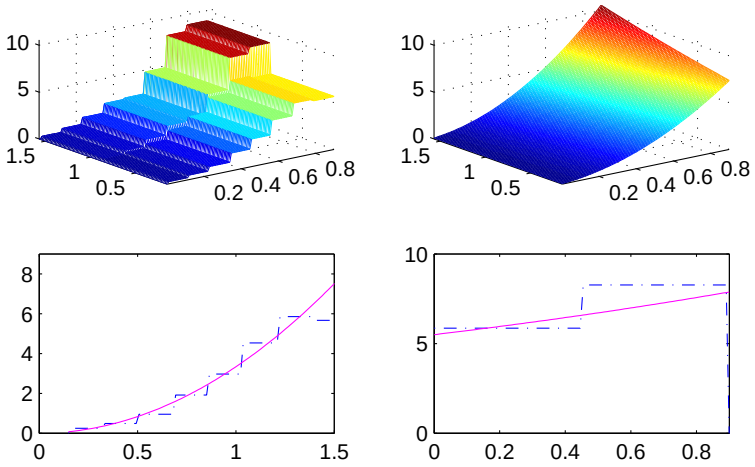

FIGURE 3. Case (PH) Estimated (top left) and true (top right) conditional hazard rates and example of sections (bottom) for a fixed value of x (left) or y (right).

where we recall that f denotes the density of μ w.r.t. the Lebesgue measure on A. We shall use the following sets:

$$\hat{\Gamma}_m = \{\min \mathrm{Sp}(G_m) \geq \max(\hat{f}_0/3, n^{-1/2})\}, \quad \hat{\Gamma} := \bigcap_{m \in \mathcal{M}_n} \hat{\Gamma}_m,$$

$$\Delta := \left\{\forall h \in \mathcal{S}_n : \left|\frac{\|h\|_n^2}{\|h\|_\mu^2} - 1\right| \leq \frac{1}{2}\right\}, \text{ and } \Omega := \left\{\left|\frac{\hat{f}_0}{f_0} - 1\right| \leq \frac{1}{2}\right\}. \tag{18}$$

For $m \in \mathcal{M}_n$, we denote by α_m the orthogonal projection on S_m of α restricted to A. The following decomposition holds:

$$\begin{aligned}\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\mathbb{1}(A)\|_\mu^2) \leq 2\|\alpha\mathbb{1}(A) - \alpha_m\|_\mu^2 &+ 2\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta \cap \Omega)) \\ &+ 2\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}((\Delta \cap \Omega)^\complement)). \end{aligned} \tag{19}$$

The last term is bounded via the following Proposition:

Proposition 2. *Under the Assumptions of Theorem 1,*

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}((\Delta \cap \Omega)^\complement)) \leq C_1/n, \tag{20}$$

where C_1 is a constant depending on τ, ϕ_1, ϕ_2, $\|\alpha\|_{\infty,A}$, f_0, f_1.

To study the term $\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta \cap \Omega))$, two preliminary remarks have to be made. The first one is the following Lemma:

Lemma 1. *Under the Assumptions of Theorem 1, the following embedding holds: for $n \geq 4/f_0^2$, we have*

$$\Delta \cap \Omega \subset \hat{\Gamma} \cap \Omega.$$

As a consequence, for all $m \in \mathcal{M}_n$, the matrices G_m are invertible on $\Delta \cap \Omega$. The second remark is the following useful decomposition. Let us define

$$\begin{aligned}\nu_n(h) &= \frac{1}{n}\sum_{i=1}^n \Big(\int_0^\tau h(X_i, z) dN^i(z) - \int_0^\tau h(X_i, z)\alpha(X_i, z) Y^i(z) dz \Big) \\ &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau h(X_i, z) dM^i(z), \end{aligned} \tag{21}$$

where we use the Doob-Meyer decomposition. For any $h_1, h_2 \in (L^2 \cap L^\infty)(A)$, we have

$$\begin{aligned}\gamma_n(h_1) - \gamma_n(h_2) &= \|h_1 - h_2\|_n^2 + 2\langle h_1 - h_2, h_2\rangle_n - \frac{2}{n}\sum_{i=1}^n \int_0^\tau (h_1 - h_2)(X_i, z) dN^i(z) \\ &= \|h_1 - h_2\|_n^2 + 2\langle h_1 - h_2, h_2 - \alpha\rangle_n - 2\nu_n(h_1 - h_2) \\ &= \|h_1 - h_2\|_n^2 + 2\langle h_1 - h_2, h_2 - \alpha\mathbb{1}(A)\rangle_n - 2\nu_n(h_1 - h_2), \end{aligned} \tag{22}$$

where the indicator $\mathbb{1}(A)$ is inserted because all other functions in the product are A-supported. Let us assume that $n \geq 4/f_0^2$. Now, on $\Delta \cap \Omega$, we have thanks to Lemma 1, by the definition of $\hat{m}$, that

$$\gamma_n(\hat{\alpha}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(\alpha_m) + \mathrm{pen}(m) \quad \forall m \in \mathcal{M}_n.$$

It follows from (22) and from the fact that $2xy \leq x^2/\theta + \theta y^2$ for any $x, y, \theta > 0$ that, on $\Delta \cap \Omega$,

$$\begin{aligned}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 &\leq 2\langle \hat{\alpha}_{\hat{m}} - \alpha_m, \alpha\mathbb{1}(A) - \alpha_m\rangle_n + \mathrm{pen}(m) + 2\nu_n(\hat{\alpha}_{\hat{m}} - \alpha_m) - \mathrm{pen}(\hat{m}) \\ &\leq \frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 + 4\|\alpha\mathbb{1}(A) - \alpha_m\|_n^2 + \mathrm{pen}(m) \\ &\quad + \frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 + 4 \sup_{h \in B^\mu_{m,\hat{m}}(0,1)} \nu_n^2(h) - \mathrm{pen}(\hat{m}), \end{aligned}$$

where $B^{\mu}_{m,m'}(0,1) := \{h \in S_m + S_{m'} : \|h\|_\mu \leq 1\}$. Now, we need to introduce a centering factor denoted by $p(m,m')$, related to the supremum of the empirical process $\nu_n(h)$:

Proposition 3. *Grant the assumptions of Theorem 1. There exists a numerical constant $\kappa > 0$ such that the following holds. If*

$$p(m,m') = \kappa(1 + \|\alpha\|_{\infty,A})\frac{D_m + D_{m'}}{n},$$

then

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\Big(\sup_{h \in B^{\mu}_{m,m'}(0,1)} (\nu_n^2(h) - p(m,m'))_+ \mathbb{1}(\Delta)\Big) \leq \frac{C_2}{n},$$

for n large enough, where C_2 is a constant depending on f_0, $\|\alpha\|_{\infty,A}$ and the chosen basis (see Section 2.4).

The proof of Proposition 3 is given in Section 6.4 below. It yields

$$\begin{aligned}\frac{3}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_n^2 \leq 4\|\alpha\mathbb{1}(A) - \alpha_m\|_n^2 + \mathrm{pen}(m) + \frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \\ + 4\Big(\sup_{h \in B^{\mu}_{m,\hat{m}}(0,1)} \nu_n^2(h) - p(m,\hat{m})\Big)_+ + 4p(m,\hat{m}) - \mathrm{pen}(\hat{m}).\end{aligned}$$

Now, let fix $K_0 \geq 4\kappa$, so that

$$4p(m,m') \leq \mathrm{pen}(m) + \mathrm{pen}(m') \quad \forall m, m',$$

and use the definition of Δ. We obtain on $\Delta \cap \Omega$:

$$\begin{aligned}\frac{3}{8}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \leq 4\|\alpha\mathbb{1}(A) - \alpha_m\|_n^2 + 2\mathrm{pen}(m) \\ + \frac{1}{4}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 + 4\sum_{m' \in \mathcal{M}_n} \Big(\sup_{h \in B^{\mu}_{m,m'}(0,1)} \nu_n^2(h) - p(m,m')\Big)_+ \end{aligned} \tag{23}$$

and thus on $\Delta \cap \Omega$:

$$\frac{1}{8}\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \leq 4\|\alpha\mathbb{1}(A) - \alpha_m\|_n^2 + 2\mathrm{pen}(m) + 4\sum_{m' \in \mathcal{M}_n} \Big(\sup_{h \in B^{\mu}_{m,m'}(0,1)} \nu_n^2(h) - p(m,m')\Big)_+.$$

Now, Proposition 3 entails:

$$\frac{1}{8}\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta \cap \Omega)) \leq 4\|\alpha\mathbb{1}(A) - \alpha_m\|_\mu^2 + 2\mathrm{pen}(m) + \frac{C_2}{n}. \tag{24}$$

Gathering (19), (20) and (24), we obtain that, for $n \geq 4/f_0^2$,

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\mathbb{1}(A)\|_\mu^2) \leq 2\|\alpha_m - \alpha\mathbb{1}(A)\|_\mu^2 + 16\Big(4\|\alpha\mathbb{1}(A) - \alpha_m\|_\mu^2 + 2\mathrm{pen}(m) + \frac{C_2}{n}\Big) + \frac{2C_1}{n}$$

for any $m \in \mathcal{M}_n$. On the other hand, if $n \leq 4/f_0^2$, then $1/n \geq f_0^2/4$ and it is easy to see that Lemma 3 (see below) entails $\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha\mathbb{1}(A)\|_\mu^2) \leq C/n$ where C is a constant depending on C_B from Lemma 3, f_0 and $\|\alpha\mathbb{1}(A)\|_\mu^2$. This concludes the proof of Theorem 1. □

5.2. **Proof of Corollary 1.** To control the bias term, we use Lemma 6, see Appendix A.2, that gives the approximation result allowing to derive the rate of convergence. If we choose S_m as one of the finite linear span considered in Section A.2, we can apply Lemma 6 to the function α_A, the restriction of α to A. Since α_m has been defined as the orthogonal projection of α_A on S_m, we get using $(\mathcal{A}1)$ and $(\mathcal{A}4)$:

$$\|\alpha\mathbb{1}(A) - \alpha_m\|_\mu \leq f_1\|\alpha - \alpha_m\|_A \leq C_3[D_{m_1}^{-\beta_1} + D_{m_2}^{-\beta_2}]$$

where C_3 depends on the Besov norm of α and on f_1. Now, according to Theorem 1 and $(\mathcal{A}2)$, we obtain:

$$\mathbb{E}\|\hat{\alpha}_{\hat{m}} - \alpha\|_A^2 \leq f_0^{-1}\mathbb{E}\left(\|\hat{\alpha}_{\hat{m}} - \alpha\|_\mu^2\right) \leq C_4 \inf_{m\in\mathcal{M}_n}\left\{D_{m_1}^{-2\beta_1} + D_{m_2}^{-2\beta_2} + \frac{D_{m_1}D_{m_2}}{n}\right\}$$

where C_4 depends on the Besov norm of α, and on f_0, f_1, ϕ_1, ϕ_2 and τ. In particular, if $m^* = (m_1^*, m_2^*)$ is such that

$$D_{m_1^*} = \lfloor n^{\frac{\beta_2}{\beta_1+\beta_2+2\beta_1\beta_2}} \rfloor \text{ and } D_{m_2^*} = \lfloor (D_{m_1^*})^{\frac{\beta_1}{\beta_2}} \rfloor$$

then

$$\mathbb{E}\|\hat{\alpha}_{\hat{m}} - \alpha\|_A^2 \leq 2C_4\Big(D_{m_1^*}^{-2\beta_1} + \frac{D_{m_1^*}^{1+\beta_1/\beta_2}}{n}\Big) \leq 4n^{-\frac{2\beta_1\beta_2}{\beta_1+\beta_2+2\beta_1\beta_2}} = 4C_4 n^{-\frac{2\bar{\beta}}{2\bar{\beta}+2}},$$

where we recall that the harmonic mean of β_1 and β_2 is $\bar{\beta} = 2\beta_1\beta_2/(\beta_1+\beta_2)$. The condition $D_{m_1} \leq \sqrt{n}/\log n$ allows this choice of m^* only if $\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2) < 1/2$ i.e. if $\beta_1 - \beta_2 + 2\beta_1\beta_2 > 0$. In the same manner, the condition $\beta_2 - \beta_1 + 2\beta_1\beta_2 > 0$ must be satisfied. Both conditions hold if $\beta_1 > 1/2$ and $\beta_2 > 1/2$. □

5.3. **Proof of Theorem 2.** The proof follows the line of the proof of Theorem 2.2 p. 67 in Lacour (2007b), so we only give a sketch of proof. Let us define

$$\Lambda = \left\{\left|\frac{\|\hat{\alpha}_{m^*}\|_\infty}{\|\alpha\|_{\infty,A}} - 1\right| < \frac{1}{2}\right\},$$

and recall that Δ and Ω are given by (18). Then we decompose the risk of $\hat{\alpha}_{\hat{\hat{m}}}$ as follows:

$$\begin{aligned}\mathbb{E}(\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha\mathbb{1}(A)\|_\mu^2) &= \mathbb{E}(\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha\mathbb{1}(A)\|_\mu^2\mathbb{1}(\Lambda\cap\Delta\cap\Omega)) \\ &\quad + \mathbb{E}(\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha\mathbb{1}(A)\|_\mu^2\mathbb{1}((\Lambda\cap\Delta\cap\Omega)^\complement)).\end{aligned}$$

The study of the term $\mathbb{E}(\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha\mathbb{1}(A)\|_\mu^2\mathbb{1}(\Lambda\cap\Delta\cap\Omega))$ is very similar to the study of its analogous in the proof of Theorem 1, by using that, on Λ,

$$\frac{1}{2}\mathrm{pen}(m) \leq \frac{K_0}{K_1}\widehat{\mathrm{pen}}(m) \leq \frac{3}{2}\mathrm{pen}(m). \tag{25}$$

Thus, the algebra starts with $\widehat{\mathrm{pen}}(m)$ instead of $\mathrm{pen}(m)$, and on Λ, it is proportional to $\mathrm{pen}(m)$ thanks to (25). At the end, only constant multiplicative factors are changed. In other words, taking $K_1 = 2K_0$, (24) is simply replaced by

$$\frac{1}{8}\mathbb{E}(\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha_m\|_\mu^2\mathbb{1}(\Delta\cap\Omega\cap\Lambda)) \leq 4\|\alpha\mathbb{1}(A) - \alpha_m\|_\mu^2 + 4\mathrm{pen}(m) + \frac{C_2}{n}. \tag{26}$$

The conclusion follows from the following Lemma, which is proven in Section 6.5:

Lemma 2. *Under the assumptions of Theorem 2,*

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha \mathbb{1}(A)\|_\mu^2 \mathbb{1}((\Lambda \cap \Delta \cap \Omega)^\complement)) \leq C_R/n,$$

where C_R depends on $\phi_1, \phi_2, \tau, f_0, f_1$ and $\|\alpha\|_{\infty,A}$.

This ends the proof of Theorem 2. □

5.4. **Proof of Theorem 3.** In order to prove Theorem 3, we use the following theorem from Tsybakov (2003), which is a standard tool for the proof of such a lower bound. We say that ∂ is a *semi-distance* on some set Θ if it is symmetric and if it satisfies the triangle inequality and $\partial(\theta, \theta) = 0$ for any $\theta \in \Theta$. We consider $K(P, Q) := \int \log(dP/dQ)dP$ the Kullback-Leibler divergence between probability measures P and Q such that $P \ll Q$.

Theorem (Tsybakov (2003))**.** *Let (Θ, ∂) be a set endowed with a semi-distance ∂. We suppose that $\{P_\theta : \theta \in \Theta\}$ is a family of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$ and that $v > 0$. If there exist $\{\theta_0, \dots, \theta_M\} \subset \Theta$, with $M \geq 2$, such that*

(1) $\partial(\theta_j, \theta_k) \geq 2v \quad \forall\, 0 \leq j < k \leq M$
(2) $P_{\theta_j} \ll P_{\theta_0} \quad \forall\, 1 \leq j \leq M$,
(3) $\frac{1}{M}\sum_{j=1}^M K(P_{\theta_j}, P_{\theta_0}) \leq a \log(M)$ *for some* $a \in (0, 1/8)$,

then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} E_\theta[(v^{-1}\partial(\hat{\theta}, \theta))^2] \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2a - 2\sqrt{\frac{a}{\log(M)}}\right),$$

where the infimum is taken among all estimators.

In this proof, we denote by $\mathbb{P}_\alpha$ the distribution of $(X, N(z), Y(z); z \leq \tau)$ when the intensity of N is α and by $\mathbb{P}_\alpha^n$ the distribution of the n-sample $(X_i, N^i(z), Y^i(z); z \leq \tau, i = 1, \dots, n)$.

We construct a family of functions $\{\alpha_0, \dots, \alpha_M\}$ that satisfies points (1)–(3). We use the notation $|A|$ for the area of the rectangle A (or the length of an interval) and $\#(R)$ denotes the cardinality of a set R. Let $\alpha_0(x, t) = |B|^{-1}\mathbb{1}(t \in B)$ where B is a compact set such that $A = [0, 1] \times [0, \tau] \subset B \times B$ and $|B| \geq 2|A|^{1/2}/L$. As a consequence, we have $\alpha_0(x, t) > 0$ for $(x, t) \in A$ and $\|\alpha_0\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} = \|\alpha_0\|_A + |\alpha_0|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} \leq L/2$ since $|\alpha_0|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} = 0$, see (10). We shall denote for short $a_0 = |B|^{-1}$ in the following. Let ψ be a very regular wavelet with compact support (the Daubechies's wavelet for instance), and for $j = (j_1, j_2) \in \mathbb{Z}^2$ and $k = (k_1, k_2) \in \mathbb{Z}^2$, let us consider

$$\psi_{j,k}(x, t) = \tau^{-1/2} 2^{(j_1+j_2)/2} \psi(2^{j_1} t/\tau - k_1)\psi(2^{j_2} x - k_2),$$

so that $\|\psi_{j,k}\|_A = 1$. Let $S_{j,k}$ stands for the support of $\psi_{j,k}$. We consider the maximal set $R_j \subset \mathbb{Z}^2$ such that

$$S_{j,k} \subset A, \forall k \in R_j \text{ and } S_{j,k} \cap S_{j,k'} = \emptyset, \forall k, k' \in R_j, k \neq k'. \tag{27}$$

The cardinality of R_j satisfies $\#(R_j) = c2^{j_1+j_2}$, where c is a positive constant that depends on τ and on the support of ψ only. Consider the set $\Omega_j = \{0, 1\}^{\#(R_j)}$ and define for any $\omega = (\omega_k) \in \Omega_j$

$$\alpha(\cdot; \omega) := \alpha_0 + \sqrt{\frac{b}{n}} \sum_{k \in R_j} \omega_k \psi_{j,k},$$

where $b > 0$ is some constant to be chosen below. In view of (27) we have

$$\|\alpha(\cdot;\omega) - \alpha(\cdot;\omega')\|_A^2 = \frac{b\rho(\omega,\omega')}{n}$$

where

$$\rho(\omega,\omega') := \sum_{k\in R_j} \mathbb{1}(\omega_k \neq \omega'_k)$$

is the Hamming distance on Ω_j. Using a result of Varshamov-Gilbert - see Tsybakov (2003) - we can find a subset $\{\omega^{(0)},\dots,\omega^{(M_j)}\}$ of Ω_j such that

$$\omega^{(0)} = (0,\dots,0), \quad \rho(\omega^{(p)},\omega^{(q)}) \geq \#(R_j)/8$$

for any $0 \leq p < q \leq M_j$, where $M_j \geq 2^{\#(R_j)/8}$. We consider the family $\mathcal{A}_j = \{\alpha_0,\dots,\alpha_{M_j}\}$ where $\alpha_p = \alpha(\cdot,\omega^{(p)})$. This family satisfies for any $0 \leq p < q \leq M_j$

$$\|\alpha_p - \alpha_q\|_A \geq \Big(\frac{b\#(R_j)}{8n}\Big)^{1/2} = 2v_j$$

for $v_j := \sqrt{b\#(R_j)/(32n)}$. This proves point (1). Now, let us gather here some properties for this family of functions. We have

$$\|\alpha(\cdot;\omega) - \alpha_0\|_{\infty,A} \leq \sqrt{\frac{b2^{(j_1+j_2)}}{\tau n}}\|\psi\|_\infty^2 \leq a_0/3$$

and consequently $\alpha(x,t;\omega) \geq 2a_0/3 > 0$ for any $(x,t) \in A$ and $\omega \in \Omega_j$ whenever

$$\Big(\frac{b2^{j_1+j_2}}{\tau n}\Big)^{1/2} \leq \frac{a_0}{3\|\psi\|_\infty^2}. \tag{28}$$

Using the Bernstein's estimate from Hochmuth (2002) (see Theorem 3.5 p.194), we have for ψ smooth enough that

$$\|\sum_{k\in R_j} \omega_k\psi_{j,k}\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} \leq c_\tau(2^{j_1\beta_1} + 2^{j_2\beta_2})\|\sum_{k\in R_j} \omega_k\psi_{j,k}\|_A \leq c_{\tau,\psi}(2^{j_1\beta_1} + 2^{j_2\beta_2})(2^{j_1+j_2})^{1/2},$$

where $c_{\tau,\psi}$ is a constant that depends on τ and ψ. Note that the Bernstein's estimate from Hochmuth (2002) is stated on the space $\mathbb{L}^2([0,1]^2)$ while we consider here $\mathbb{L}^2([0,1]\times[0,\tau])$. An obvious (but tedious) modification of the proof of Hochmuth (it suffices to change the scaling of the moduli of continuity $\omega_{r_i,i}$ herein) allows to show that the Bernstein's estimate is the same as for $\mathbb{L}^2([0,1]^2)$, up to a multiplicative constant that depends on τ. Hence, if

$$\frac{c_{\tau,\psi}(2^{j_1\beta_1} + 2^{j_2\beta_2})(2^{j_1+j_2})^{1/2}}{\sqrt{n}} \leq \frac{L}{2\sqrt{b}}, \tag{29}$$

we have $\|\alpha(\cdot;\omega)\|_{B^{\boldsymbol{\beta}}_{2,\infty}(A)} \leq L$, so $\alpha(\cdot;\omega) \in B^{\boldsymbol{\beta}}_{2,\infty}(A,L)$ for any $\omega \in \Omega_j$. This proves that $\mathcal{A}_j \subset B^{\boldsymbol{\beta}}_{2,\infty}(A,L)$.

Points (2) and (3) are derived using Jacod's formula (see Andersen et al. (1993)). Indeed, we can prove that the log-likelihood $\ell(\alpha,\alpha_0) := \log(d\mathbb{P}_\alpha/d\mathbb{P}_{\alpha_0})$ of N writes

$$\ell(\alpha,\alpha_0) = \int_0^\tau (\log\alpha(X,t) - \log\alpha_0(X,t))dN(t) - \int_0^\tau (\alpha(X,t) - \alpha_0(X,t))Y(t)dt.$$

For any $\alpha \in \mathcal{A}_j$, we have $\|\alpha - \alpha_0\|_{\infty,A} \leq a_0/3 \leq \alpha(x,t)/2$ for any $(x,t) \in A$. The Doob-Meyer decomposition allows to write that, under $\mathbb{P}_{\alpha_0}$:

$$\ell(\alpha,\alpha_0) = \int_0^\tau \Big(\Phi_{1/\alpha(X,t)}(\alpha(X,t)-\alpha_0(X,t)) - (\alpha(X,t)-\alpha_0(X,t))\Big)Y(t)dt + \int_0^\tau (\log\alpha(X,t) - \log\alpha_0(X,t))dM(t)$$

where $\Phi_a(x) := -\log(1-ax)/a$ for $a > 0$ and $x < 1/a$. But since $\Phi_a(x) \leq x + ax^2$ for any $x \leq 1/(2a)$, we obtain

$$\ell(\alpha,\alpha_0) \leq \frac{3}{2a_0}\int_0^\tau (\alpha(t,X)-\alpha_0(t,X))^2 Y(t)dt + \int_0^\tau (\log\alpha_0(t,X) - \log\alpha(t,X))dM(t)$$

which gives by integration with respect to $\mathbb{P}_\alpha$

$$K(\mathbb{P}_\alpha,\mathbb{P}_{\alpha_0}) \leq \frac{3\|\alpha-\alpha_0\|_\mu^2}{2a_0} \leq \frac{3f_1\|\alpha-\alpha_0\|_A^2}{2a_0} \leq \frac{3bf_1\#(R_j)}{2na_0},$$

for any $\alpha \in \mathcal{A}_j$. Since the counting processes $(N^1,\ldots,N^n)$ are independent, we have $K(\mathbb{P}_\alpha^n,\mathbb{P}_{\alpha_0}^n) = nK(\mathbb{P}_\alpha,\mathbb{P}_{\alpha_0})$ and

$$\frac{1}{M}\sum_{p=0}^M K(\mathbb{P}_{\alpha_p}^n,\mathbb{P}_{\alpha_0}^n) \leq \frac{3bf_1\#(R_j)}{2a_0} \leq a\log M_j$$

with

$$a := 12bf_1/(a_0\log 2).$$

So, we take b small enough, so that $a < 1/8$ (this is the only constraint on b) and point (3) is met in Tsybakov (2003)'s theorem. It only remains to choose the levels j_1 and j_2 so that (28) and (29) holds, and to compute the corresponding v_j. We take $j = (j_1,j_2)$ such that

$$c_1/2 \leq 2^{j_1} n^{-\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq c_1 \text{ and } c_2/2 \leq 2^{j_2} n^{-\beta_1/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq c_2$$

where c_1 and c_2 are positive constants satisfying $c_{\tau,\psi}(c_1^{\beta_1}+c_2^{\beta_2})\sqrt{c_1c_2} \leq L/(2\sqrt{b})$. For this choice, $2^{j_1+j_2}/n \leq c_1c_2n^{-2\bar\beta/(2\bar\beta+2)}$ so (28) holds for n large enough and (29) holds and $v_j \geq c_3 n^{-\bar\beta/(2\bar\beta+2)}$ where $c_3 = c_{\tau,\psi}\sqrt{bc_1c_2/128}$. □

6. Proof of the auxiliary results

6.1. **Proof of Proposition 1.** Let $\hat f_{m_1^*}$ and $\hat f_0$ be defined by (16), with $m_1^* = (D_{m_1},\mathcal{D}_n^{(2)})$ and $D_{m_1} = \mathcal{D}_n^{(2)} = n^{1/4}/\sqrt{\log n}$. We remark that, for all $(x,z) \in \mathbb{R}^2$,

$$\hat f_{m_1^*}(x,z) = f(x,z) + \hat f_{m_1^*}(x,z) - f(x,z) \geq f_0 - \|\hat f_{m_1^*} - f\|_{\infty,A}.$$

We deduce that $\|\hat f_{m_1^*} - f\|_{\infty,A} \geq f_0 - \hat f_0$. In the same manner, $\|\hat f_{m_1^*} - f\|_{\infty,A} \geq \hat f_0 - f_0$. Thus

$$\mathbb{P}(\Omega^{\complement}) = \mathbb{P}(|f_0 - \hat f_0| > f_0/2) \leq \mathbb{P}(\|\hat f_{m_1^*} - f\|_{\infty,A} > f_0/2).$$

Therefore, we just have to prove that $\mathbb{P}(\|\hat f_{m_1^*} - f\|_{\infty,A} > f_0/2) \leq C_0/n^k$. First, remark that $\|\hat f_{m_1^*} - f\|_{\infty,A} \leq \|\hat f_{m_1^*} - f_{m_1^*}\|_{\infty,A} + \|f_{m_1^*} - f\|_{\infty,A}$. As $f \in B_{2,\infty}^{(\tilde\beta_1,\tilde\beta_2)}(A)$ with $\bar{\tilde\beta} > 1$, the embedding theorem proved in Nikol'skii (1975) p.236 implies that f belongs to $B_{\infty,\infty}^{(\beta_1^*,\beta_2^*)}(A)$

with $\beta_1^* = \tilde{\beta}_1(1-1/\bar{\tilde{\beta}})$ and $\beta_2^* = \tilde{\beta}_2(1-1/\bar{\tilde{\beta}})$. Moreover, Nikol'skii (1975) proves that there exists a function F_{m^*} in the space S_{m^*} of trigonometric polynomials such that

$$\|F_{m^*} - f\mathbb{1}(A)\| \le C(D_{m_1^*}^{-\tilde{\beta}_1} + \mathcal{D}_n^{-\tilde{\beta}_2}) \text{ and } \|F_{m^*} - f\mathbb{1}(A)\|_\infty \le C(D_{m_1^*}^{-\beta_1^*} + \mathcal{D}_n^{-\tilde{\beta}_2^*}),$$

where C depends on the Besov norm of f on A. Then

$$\begin{aligned}
\|f_{m_1*} - f\mathbb{1}(A)\|_\infty &\le \|f_{m_1*} - F_{m^*}\|_\infty + \|F_{m^*} - f\mathbb{1}(A)\|_\infty \\
&\le \phi_0\sqrt{D_{m_1*}\mathcal{D}_n^{(2)}}\|f_{m_1*} - F_{m^*}\| + \|F_{m^*} - f\mathbb{1}(A)\|_\infty \\
&\le \phi_0\sqrt{D_{m_1*}\mathcal{D}_n^{(2)}}(\|f_{m_1*} - f\mathbb{1}(A)\| + \|f\mathbb{1}(A) - F_{m^*}\|) + \|F_{m^*} - f\mathbb{1}(A)\|_\infty \\
&\le C'[\sqrt{D_{m_1*}\mathcal{D}_n^{(2)}}(D_{m_1^*}^{-\tilde{\beta}_1} + \mathcal{D}_n^{-\tilde{\beta}_2}) + D_{m_1*}^{-\beta_1^*} + (\mathcal{D}_n^{(2)})^{-\beta_2^*}],
\end{aligned}$$

where C' depends on ϕ_0 and the Besov norm of f. But since $D_{m_1^*} = \mathcal{D}_n^{(2)} = n^{1/4}/\log(n)$, this proves that $\|f_{m_1*} - f\mathbb{1}(A)\|_\infty \to 0$ when $n \to +\infty$ as soon as $\tilde{\beta}_1 > 2$ and $\tilde{\beta}_2 > 2$. So, there is n_0 such that for any $n \ge n_0$, we have $\|f_{m_1*} - f\|_{\infty,A} \le f_0/4$ and

$$\mathbb{P}(\|\hat{f}_{m_1^*} - f\|_{\infty,A} > f_0/2) \le \mathbb{P}(\|\hat{f}_{m_1^*} - f_{m_1^*}\|_{\infty,A} > f_0/4).$$

Using $(\mathcal{M}2)$, we get

$$\|\hat{f}_{m_1^*} - f_{m_1^*}\|_{\infty,A} \le \sqrt{\phi_1\phi_2 D_{m_1^*}\mathcal{D}_n^{(2)}}\|\hat{f}_{m_1^*} - f_{m_1^*}\|.$$

Now we define

$$\vartheta_n(h) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big(h(X_i,y)Y^i(y) - \mathbb{E}\big(h(X_i,y)Y^i(y)\big)\Big)dy = \|\sqrt{h}\|_n^2 - \|\sqrt{h}\|_\mu^2. \tag{30}$$

With this notation, and recalling the definition of $\hat{f}_m$ (see Equation (15)), we have $\mathbb{E}(\hat{b}_{j,k}) = b_{j,k}$ and

$$\|\hat{f}_{m_1^*} - f_{m_1^*}\|^2 = \sum_{j,k}(\hat{b}_{j,k} - b_{j,k})^2 = \sum_{j,k}\vartheta_n^2(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*}),$$

thus:

$$\begin{aligned}
\mathbb{P}(\|\hat{f}_{m_1^*} - f\|_{\infty,A} > f_0/2) &\le \mathbb{P}\Big(\sum_{j,k}\vartheta_n^2(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*}) \ge \frac{f_0^2}{16\phi_1\phi_2 D_{m_1^*}\mathcal{D}_n^{(2)}}\Big) \\
&\le \sum_{j,k}\mathbb{P}\Big(|\vartheta_n(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*})| \ge \frac{f_0}{4\sqrt{\phi_1\phi_2}D_{m_1^*}\mathcal{D}_n^{(2)}}\Big).
\end{aligned}$$

Note that $\vartheta_n(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*}) = \frac{1}{n}\sum_1^n(U_i^{j,k} - \mathbb{E}(U_i^{j,k}))$, where $U_i^{j,k} = \varphi_j(X_i)\int_0^\tau \psi_k(y)Y^i(y)dy$ are i.i.d. random variables. We apply the Bernstein inequality to to the sum of the random variables $U_i^{j,k}$. We have

$$|U_i^{j,k}| \le \|\varphi_j\|_\infty \int_0^\tau |\psi_k(y)|dy \le \|\varphi_j\|_\infty\Big(\tau\int_0^\tau \psi_k^2(y)dy\Big)^{1/2} \le \sqrt{\tau\phi_1 D_{m_1^*}} := c$$

and $\mathbb{E}[(U_i^{j,k})^2] \leq \tau f_1 =: v^2$, so the Bernstein inequality gives

$$\mathbb{P}\Big(|\vartheta_n(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*})| \geq x\Big) \leq 2\exp\Big(-\frac{nx^2}{2(v^2+cx/3)}\Big)$$

with $x = f_0/(4\sqrt{\phi_1\phi_2}D_{m_1^*}\mathcal{D}_n^{(2)})$ and v and c defined above. This entails:

$$\mathbb{P}\Big(|\vartheta_n(\varphi_j^{m_1^*} \otimes \psi_k^{m_1^*})| \geq \frac{f_0}{4\sqrt{\phi_1\phi_2}D_{m_1^*}\mathcal{D}_n^{(2)}}\Big) \leq 2\exp\Big(-\frac{Cn}{(D_{m_1^*}\mathcal{D}_n^{(2)})^2}\Big),$$

where C is a constant depending on $f_0, f_1, \tau, \phi_1, \phi_2$, and since $D_{m_1^*} = \mathcal{D}_n^{(2)} = n^{1/4}/\sqrt{\log(n)}$ we obtain:

$$\mathbb{P}(\Omega^{\complement}) \leq 2\sqrt{n}\exp\big(-C(\log n)^2\big) \leq \frac{C_0}{n^k},$$

where C_0 is a constant depending on k, f_0, ϕ_1, ϕ_2, τ and f_1. This concludes the proof of Proposition 1. □

6.2. **Proof of Proposition 2.**

6.2.1. *Proof of Proposition 2.* One can write

$$\begin{aligned}\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}((\Delta\cap\Omega)^{\complement})) &\leq \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Delta^{\complement})) + \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}(\Omega^{\complement}))\\ &\leq f_1[\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|^2 \mathbb{1}(\Delta^{\complement})) + \mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|^2 \mathbb{1}(\Omega^{\complement}))]\end{aligned}$$

using $(\mathcal{A}1)$ and $(\mathcal{A}4)$. This yields

$$\begin{aligned}\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}((\Delta\cap\Omega)^{\complement})) \leq 2f_1\big[&\mathbb{E}^{1/2}(\|\hat{\alpha}_{\hat{m}}\|^4)(\mathbb{P}^{1/2}(\Delta^{\complement}) + \mathbb{P}^{1/2}(\Omega^{\complement}))\\ &+ \|\alpha\|_A^2(\mathbb{P}(\Omega^{\complement}) + \mathbb{P}(\Delta^{\complement}))\big].\end{aligned}$$

Now, $(\mathcal{A}5)$ with $k=7$ ensures that $\mathbb{P}(\Omega^{\complement}) \leq C_0/n^7$ for any $n \geq n_0$. We need the following Lemmas:

Lemma 3. *Under the assumptions of Theorem 1, $\mathbb{E}(\|\hat{\alpha}_{\hat{m}}\|^4) \leq C_B n^5$, where C_B is a constant depending on ϕ_1, ϕ_2, τ and $\|\alpha\|_{\infty,A}$.*

Lemma 4. *Under the assumptions of Theorem 1, we have $\mathbb{P}(\Delta^{\complement}) \leq C_k^{(\Delta)}/n^k$ for any $k \geq 1$, where $C_k^{(\Delta)}$ is a constant depending on k, on the basis, and on f_0, f_1.*

Using Lemmas 3 and 4 and Assumption $(\mathcal{A}5)$, we get

$$\mathbb{E}(\|\hat{\alpha}_{\hat{m}} - \alpha_m\|_\mu^2 \mathbb{1}((\Delta\cap\Omega)^{\complement})) \leq C_1/n, \tag{31}$$

where C_1 is a constant depending on τ, ϕ_1, ϕ_2, $\|\alpha\|_{\infty,A}$, f_0, f_1. This concludes the proof of Proposition 2. □

6.2.2. *Proof of Lemma 3.* Note that $\hat{\alpha}_{\hat{m}}$ is either 0 or $\operatorname{argmin}_{t\in S_{\hat{m}}}\gamma_n(t)$. Let us denote for short $\varphi_j := \varphi_j^{\hat{m}}$ and $\psi_k := \psi_k^{\hat{m}}$. In the second case, $\min \mathrm{Sp}(G_{\hat{m}}) \geq \max(\hat{f}_0/3, n^{-1/2})$, so

$$\begin{aligned}
\|\hat{\alpha}_{\hat{m}}\|^2 &= \sum_{j,k}(\hat{a}_{j,k}^{\hat{m}})^2 = \|A_{\hat{m}}\|^2 = \|G_{\hat{m}}^{-1}\Upsilon_{\hat{m}}\|^2 \\
&\leq (\min \mathrm{Sp}(G_{\hat{m}}))^{-2}\|\Upsilon_{\hat{m}}\|^2 \leq \min(9/\hat{f}_0^2, n)\sum_{j,k}\Big(\frac{1}{n}\sum_{i=1}^n \varphi_j(X_i)\int_0^\tau \psi_k(z)dN^i(z)\Big)^2 \\
&\leq \min(9/\hat{f}_0^2, n)\frac{1}{n}\sum_{i=1}^n\sum_j \varphi_j^2(X_i)\sum_k\Big(\int_0^\tau \psi_k(z)dN^i(z)\Big)^2 \\
&\leq \min(9/\hat{f}_0^2, n)\phi_1\mathcal{D}_n^{(1)}\frac{1}{n}\sum_{i=1}^n\sum_k\Big(\mathbb{1}_{A_1}(X_i)\int_0^\tau \psi_k(z)dN^i(z)\Big)^2,
\end{aligned}$$

and

$$\begin{aligned}
\|\hat{\alpha}_{\hat{m}}\|^4 &\leq n^2\phi_1^2(\mathcal{D}_n^{(1)})^2\frac{1}{n}\sum_{i=1}^n\Big(\sum_k\Big(\mathbb{1}_{A_1}(X_i)\int_0^\tau \psi_k(z)dN^i(z)\Big)^2\Big)^2 \\
&\leq n^2\phi_1^2(\mathcal{D}_n^{(1)})^2\mathcal{D}_n^{(2)}\frac{1}{n}\sum_{i=1}^n\sum_k\Big(\mathbb{1}_{A_1}(X_i)\int_0^\tau \psi_k(z)dN^i(z)\Big)^4. \qquad (32)
\end{aligned}$$

Now, we have:

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n\sum_k\Big(\mathbb{1}_{A_1}(X_i)\int_0^\tau \psi_k(z)dN^i(z)\Big)^4\Big) \\
&\leq \frac{2^3}{n}\sum_{i=1}^n\sum_k\mathbb{E}\Big(\Big(\mathbb{1}_{A_1}(X_i)\int_0^\tau \psi_k(z)\alpha(X^i,z)Y^i(z)dz\Big)^4\Big) \qquad (33) \\
&\quad + \frac{2^3}{n}\sum_{i=1}^n\sum_k\mathbb{E}\Big(\Big(\int_0^\tau \psi_k(z)dM^i(z)\Big)^4\Big).
\end{aligned}$$

Using the Bürkholder Inequality as recalled in Liptser and Shiryayev (1989) p.75, and the fact that the quadratic variation process of each M^i is N^i $(i = 1, \ldots, n)$, we know that there exists a universal constant κ_b such that:

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n\sum_k\Big(\int_0^\tau \psi_k(z)dM^i(z)\Big)^4\Big) \leq \kappa_b\frac{1}{n}\sum_{i=1}^n\sum_k\mathbb{E}\Big(\Big(\int_0^\tau \psi_k^2(z)dN^i(z)\Big)^2\Big) \\
&\leq \kappa_b\frac{1}{n}\sum_{i=1}^n\sum_k\mathbb{E}\Big(N^i(\tau)\sum_{s:\Delta N^i(s)\neq 0}\psi_k^4(s)\Big) \leq \kappa_b\frac{1}{n}\sum_{i=1}^n\mathbb{E}\Big(N^i(\tau)\sum_{s:\Delta N^i(s)\neq 0}\sum_k\psi_k^4(s)\Big) \\
&\leq \kappa_b\phi_2^2(\mathcal{D}_n^{(2)})^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}\Big(N^i(\tau)\sum_{s:\Delta N^i(s)\neq 0}1\Big) \leq \kappa_b\phi_2^2(\mathcal{D}_n^{(2)})^2\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(N^i(\tau))^2], \qquad (34)
\end{aligned}$$

where we used Assumption $(\mathcal{M}1)$. Using Assumptions $(\mathcal{A}3)$ and $(\mathcal{A}4)$ we have

$$[N^1(\tau)]^2 = \Big(M^1(\tau) + \int_0^\tau \alpha(X_1,z)Y^1(z)dz\Big)^2 \leq 2(M^1(\tau))^2 + 2\Big(\int_0^\tau \alpha(X_1,z)Y^1(z)dz\Big)^2$$

Therefore, as $\mathbb{E}[(M^1(\tau))^2] = \mathbb{E}\int_0^\tau \alpha(X_1,z)Y^1(z)dz \le \tau\|\alpha\|_{\infty,A}$, we find

$$\mathbb{E}[(N^1(\tau))^2] \le 2\tau\|\alpha\|_{\infty,A} + 2(\tau\|\alpha\|_{\infty,A})^2. \tag{35}$$

Combining (33), (34) and (35) gives

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n\sum_k\Big(\int_0^\tau \psi_k(z)dN^i(z)\Big)^4\Big)\\
&\quad\le 8\kappa_b\phi_2^2(\mathcal{D}_n^{(2)})^2\mathbb{E}[(N^1(\tau))^2] + 8\sum_k\mathbb{E}\Big(\Big(\mathbb{1}_{A_1}(X)\int_0^\tau\psi_k(z)\alpha(X,z)Y(z)dz\Big)^4\Big)\Big)\\
&\quad\le 8\kappa_b\phi_2^2(\mathcal{D}_n^{(2)})^2\mathbb{E}[(N^1(\tau))^2] + 8\|\alpha\|_{\infty,A}^4\tau^2\sum_k\Big(\int_0^\tau\psi_k^2(z)dz\Big)^2\\
&\quad\le 8\kappa_b\phi_2^2(\mathcal{D}_n^{(2)})^2\mathbb{E}[(N^1(\tau))^2] + 8\|\alpha\|_{\infty,A}^4\tau^2\mathcal{D}_n^{(2)}. \qquad (36)
\end{aligned}$$

Then we have, by inserting (36) in (32),

$$\begin{aligned}
\mathbb{E}(\|\hat\alpha_{\hat m}\|^4) &\le (\phi_1 n\mathcal{D}_n^{(1)})^2\mathcal{D}_n^{(2)}\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n\sum_k\Big(\int_0^\tau\psi_k(z)dN^i(z)\Big)^4\Big)\\
&\le C_B n^2(\mathcal{D}_n^{(1)})^2(\mathcal{D}_n^{(2)})^3 \le C_B n^{4.5} \le C_B n^5,
\end{aligned}$$

where C_B is a constant depending on ϕ_1, ϕ_2, τ and $\|\alpha\|_{\infty,A}$. We use here that $\mathcal{D}_n^{(i)} \le \sqrt{n}/\log(n)$ in the case of localized bases [DP], [W], [H]. Note that for basis [T], under $(\mathcal{M}1)$, the final order is smaller (namely $n^{3.25}$ instead of $n^{4.5}$). This concludes the proof of Lemma 3. □

6.2.3. *Proof of Lemma 4.* Define, for $\rho > 1$, the set

$$\Delta_\rho = \{\forall h \in \mathcal{S}_n, \Big|\|h\|_n^2/\|h\|_\mu^2 - 1\Big| \le 1 - 1/\rho\}, \tag{37}$$

where $\mathcal{S}_n$ is the set of maximal dimension of the collection. Remark that $\Delta = \Delta_2$, see (18). First we observe that:

$$\mathbb{P}(\Delta_\rho^\complement) \le \mathbb{P}\Big(\sup_{h\in B_{\mathcal{S}_n}^\mu(0,1)}|\vartheta_n(h^2)| > 1 - 1/\rho\Big)$$

where $\vartheta_n(\cdot)$ is defined by (30) and $B_{\mathcal{S}_n}^\mu(0,1) = \{t \in \mathcal{S}_n, \|t\|_\mu \le 1\}$. We denote by $(\varphi_j \otimes \psi_k)$ the L^2-orthonormal basis of $\mathcal{S}_n$. If $h(x,y) = \sum_{j,k} a_{j,k}\varphi_j(x)\psi_k(y)$, then

$$\vartheta_n(h^2) = \sum_{j,k,j',k'} a_{j,k}a_{j',k'}\vartheta_n((\varphi_j\otimes\psi_k)(\varphi_{j'}\otimes\psi_{k'})). \tag{38}$$

We obtain

$$\sup_{h\in B_{\mathcal{S}_n}^\mu(0,1)}|\vartheta_n(h^2)| \le f_0^{-1}\sup_{\sum a_{j,k}^2\le 1}\Big|\sum_{j,k,j',k'} a_{j,k}a_{j',k'}\vartheta_n((\varphi_j\otimes\psi_k)(\varphi_{j'}\otimes\psi_{k'}))\Big|. \tag{39}$$

Lemma (Baraud et al. (2001a))**.** *Let $B_{j,j'} = \|\varphi_j\varphi_{j'}\|_{\infty,A}$ and $V_{j,j'} = \|\varphi_j\varphi_{j'}\|_2$. Let, for any symmetric matrix $(A_{j,j'})$*

$$\bar\rho(A) := \sup_{\sum b_j^2\le 1}\sum_{j,j'}|b_jb_{j'}|A_{j,j'}$$

and $L(\varphi) := \max\{\bar\rho^2(V), \bar\rho(B)\}$. *Then, if* $(\mathcal{M}2)$ *is satisfied, we have* $L(\varphi) \le \phi_1(\mathcal{D}_n^{(1)})^2$, *and* $L(\varphi) \le 5\phi_1^4\mathcal{D}_n^{(1)}$, *if the basis is localized (cases [P] or [W]).*

Let us define

$$\begin{aligned} x &:= \frac{f_0^2(1-1/\rho)^2}{4f_1(\mathcal{D}_n^{(2)})^2 L(\varphi)} \text{ and} \\ \Theta &:= \Big\{\forall (j,k)\forall(j',k') \quad |\vartheta_n((\varphi_j\otimes\psi_k)(\varphi_{j'}\otimes\psi_{k'}))| \le 4\Big(B_{j,j'}x + V_{j,j'}\sqrt{2f_1x}\Big)\Big\}. \end{aligned}$$

Starting from (39), we have, on Θ:

$$\sup_{h\in B_{\mathcal{S}_n}^{\mu}(0,1)} |\vartheta_n(h^2)| \le 4f_0^{-1} \sup_{\sum a_{j,k}^2\le 1} \sum_{j,j'}\Big(\sum_{k,k'}|a_{j,k}a_{j',k'}|\Big)\Big(B_{j,j'}x + V_{j,j'}\sqrt{2f_1x}\Big).$$

Thus setting $b_j = \sum_k |a_{j,k}|$, we have $\sum_j b_j^2 \le \mathcal{D}_n^{(2)}$ and it follows that, on Θ,

$$\begin{aligned} \sup_{h\in B_{\mathcal{S}_n}^{\mu}(0,1)} |\vartheta_n(h^2)| &\le f_0^{-1}\mathcal{D}_n^{(2)} \sup_{\sum b_j^2=1}\sum_{j,j'}|b_jb_{j'}|\Big(B_{j,j'}x + V_{j,j'}\sqrt{2f_1x}\Big) \\ &\le f_0^{-1}\mathcal{D}_n^{(2)}\Big(\bar\rho(B)x + \bar\rho(V)\sqrt{2f_1x}\Big) \\ &\le (1-1/\rho)\Big(\frac{f_0(1-1/\rho)}{4\mathcal{D}_n^{(2)}f_1}\frac{\bar\rho(B)}{L(\varphi)} + \frac{1}{\sqrt{2}}\Big(\frac{\bar\rho^2(V)}{L(\varphi)}\Big)^{1/2}\Big) \\ &\le (1-1/\rho)\Big(\frac14 + \frac{1}{\sqrt2}\Big) \le (1-1/\rho). \end{aligned}$$

Therefore,

$$\mathbb{P}\Big(\sup_{t\in B_{\mathcal{S}_n}^{\mu}(0,1)} |\vartheta_n(t^2)| > 1 - \frac{1}{\rho}\Big) \le \mathbb{P}(\Theta^{\complement}).$$

Let $\phi_\lambda = \varphi_j\otimes\psi_k$ for $\lambda = (j,k)$. To bound $\mathbb{P}(\vartheta_n(\phi_\lambda\phi_{\lambda'}) \ge B_{j,j'}x + V_{j,j'}\sqrt{2f_1x})$, we will apply the Bernstein inequality given in Birgé and Massart (1998) to the i.i.d. r.v.

$$U_i^{\lambda,\lambda'} = U_i^{(j,k),(j',k')} = \varphi_j(X_i)\varphi_{j'}(X_i)\int_0^\tau \psi_k(y)\psi_{k'}(y)Y^i(y)dy. \tag{40}$$

Under $(\mathcal{A}4)$, the r.v. are bounded

$$|U_i^{\lambda,\lambda'}| \le \|\varphi_j\varphi_{j'}\|_{\infty,A}\int_0^\tau |\psi_k(y)\psi_{k'}(y)|dy \le \|\varphi_j\varphi_{j'}\|_{\infty,A} = B_{j,j'}.$$

Moreover, using $(\mathcal{A}4)$ again, we obtain:

$$(U_i^{\lambda,\lambda'})^2 \le (\varphi_j(X_i)\varphi_{j'}(X_i))^2\int_0^\tau\psi_k^2(y)dy\int_0^\tau\psi_{k'}^2(y)dy = (\varphi_j(X_i)\varphi_{j'}(X_i))^2$$

and thus

$$\mathbb{E}[(U_i^{\lambda,\lambda'})^2] \le \mathbb{E}[(\varphi_j(X_i)\varphi_{j'}(X_i))^2] \le f_1V_{j,j'}^2.$$

We get

$$\mathbb{P}(|\vartheta_n(\phi_\lambda\phi_{\lambda'})| \ge B_{j,j'}x + V_{j,j'}\sqrt{2f_1x}) \le 2e^{-nx}.$$

Given that $\mathbb{P}(\Delta_\rho^{\complement}) \leq \mathbb{P}(\Theta^{\complement}) = \sum_{\lambda,\lambda'} \mathbb{P}\Big(|\vartheta_n(\phi_\lambda \phi_{\lambda'})| > B_{j,j'}x + V_{j,j'}\sqrt{2f_1 x}\Big)$, we can write:

$$\begin{aligned}\mathbb{P}(\Delta_\rho^{\complement}) &\leq 2(\mathcal{D}_n^{(1)}\mathcal{D}_n^{(2)})^2 \exp\Big\{-\frac{nf_0^2(1-1/\rho)^2}{4f_1(\mathcal{D}_n^{(2)})^2 L(\varphi)}\Big\} \\ &\leq 2n^2 \exp\Big\{-\frac{f_0^2(1-1/\rho)^2}{4f_1}\frac{n}{(\mathcal{D}_n^{(2)})^2 L(\varphi)}\Big\}.\end{aligned}$$

Following the Lemma of Baraud et al. (2001a) above, and using Assumption $(\mathcal{M}_1)$, we have

$$(\mathcal{D}_n^{(2)})^2 L(\varphi) \leq \phi_1 (\mathcal{D}_n^{(2)}\mathcal{D}_n^{(1)})^2 \leq \phi_1 n/\log^2(n).$$

And then, for any k, there exists a constant $C_k^{(\Delta_\rho)}$ depending on k, f_0, $\|f\|_{\infty,A}$, ϕ_1 and ρ such that

$$\mathbb{P}(\Delta_\rho^{\complement}) \leq 2n^2 \exp\Big\{-\frac{f_0^2(1-1/\rho)^2}{40 f_1 \phi_1}\log^2(n)\Big\} \leq \frac{C_k^{(\Delta_\rho)}}{n^k}. \tag{41}$$

Now, if the basis is localized, the result is better. In this case, $L(\varphi) \leq 5\phi_1^4 \mathcal{D}_n^{(1)}$. Moreover, take histogram basis in (38), then all terms with $k \neq k'$ vanish and then we can take $b_j = (\sum_k a_{j,k}^2)^{1/2}$ directly. Then, as then $\sum_j b_j^2 \leq 1$, we obtain

$$\mathbb{P}(\Delta_\rho^{\complement}) \leq 2(\mathcal{D}_n^{(1)})^2 \mathcal{D}_n^{(2)} \exp\Big\{-\frac{nf_0^2(1-1/\rho)^2}{40 f_1 L(\varphi)}\Big\} \leq 2n^2 \exp\Big\{-\frac{f_0^2(1-1/\rho)^2}{40 f_1}\frac{n}{L(\varphi)}\Big\}.$$

Thus $L(\varphi) \leq 5\phi_1^4 \mathcal{D}_n^{(1)} \leq \phi_1 n/\log^2(n)$ is enough to get (41) again. The proof is easy to extend to any localized basis as $[P]$ or $[W]$ (with $\mathcal{D}_n^{(2)}$ in the bound of $\sum_j b_j^2$ replaced by $r+1$ in case $[P]$ for instance). This concludes the proof of Lemma 4. □

6.3. **Proof of Lemma 1.** Let $m \in \mathcal{M}_n$ be fixed and let ℓ be an eigenvalue of G_m. There exists $A_m \neq 0$ with coefficients $(a_\lambda)_\lambda$ such that $G_m A_m = \ell A_m$ and thus $A_m^\top G_m A_m = \ell A_m^\top A_m$. Now, take $h := \sum_\lambda a_\lambda \varphi_\lambda \in S_m$. We have $\|h\|_n^2 = A_m^\top G_m A_m$ and $\|h\|_A^2 = A_m^\top A_m$. Thus, on Δ (see (18)):

$$A_m^\top G_m A_m = \|h\|_n^2 \geq \frac{1}{2}\|h\|_\mu^2 \geq \frac{1}{2} f_0 \|h\|_A^2 = \frac{1}{2} f_0 A_m^\top A_m.$$

Therefore, on Δ, for all $m \in \mathcal{M}_n$, we have $\min \mathrm{Sp}(G_m) \geq f_0/2$. Moreover, on Ω, we have $f_0 \geq 2\hat{f}_0/3$ and $\max(\hat{f}_0/3, n^{-1/2}) = \hat{f}_0/3$, for $n \geq 4/f_0^2$. □

6.4. **Proof of Proposition 3.** Usually, in model selection (see for instance Massart (2007)), the penalty is obtained by using the so-called Talagrand's deviation inequality for the maximum of empirical processes. Since the empirical process $\nu_n(\cdot)$ (see Equation (21)) considered here is not bounded, we cannot use directly Talagrand's inequality. Using tools from van de Geer (1995), we prove Bennett and Bernstein type inequalities for $\nu_n(\cdot)$, and using a $L^2(\mu) - L^\infty$ generic chaining type of technique (see Talagrand (2005) and Baraud (2010)), we derive an uniform deviation.

Lemma 5. *For any positive δ, ϵ and for any function $h \in (L^2 \cap L^\infty)(A)$, we have the following Bennett-type deviation inequality*:

$$\mathbb{P}\big(\nu_n(h) \geq \epsilon, \|h\|_n \leq \delta\big) \leq \exp\Big(-\frac{n\delta^2\|\alpha\|_{\infty,A}}{\|h\|^2_{\infty,A}} g\Big(\frac{\epsilon\|h\|_{\infty,A}}{\|\alpha\|_{\infty,A}\delta^2}\Big)\Big)$$

where $g(x) = (1+x)\log(1+x) - x$ for any $x \geq 0$. As a consequence, we obtain the following Bernstein-type inequalities:

$$\mathbb{P}\big(\nu_n(h) \geq \epsilon, \|h\|_n \leq \delta\big) \leq \exp\Big(-\frac{n\epsilon^2/2}{\|\alpha\|_{A,\infty}\delta^2 + \epsilon\|h\|_{A,\infty}/3}\Big), \tag{42}$$

and

$$\mathbb{P}\Big(\nu_n(h) \geq \delta\sqrt{2\|\alpha\|_{\infty,A}x} + \|h\|_{\infty,A}x, \;\; \|h\|_n^2 \leq \delta^2\Big) \leq \exp(-nx). \tag{43}$$

Proof. Notice that the process

$$n\nu(h,t) := \sum_{i=1}^n \int_0^t h(X_i,z)dM^i(z) := \sum_{i=1}^n \nu(h,t)^i$$

is a locally square integrable martingale with jumps of size less than $n\|h\|_{\infty,A}$. As a consequence, Corollary 2.3. of van de Geer (1995) applies almost directly. However to introduce the empirical norm $\|h\|_n$ in the deviation inequality, we re-derive the majoration of the term

$$S_\tau := \sum_{i=1}^n \sum_{k\geq 2} \frac{a^k}{k!} \int_0^\tau |h(X_i,z)|^k dV_k^i(z),$$

where, for all $i = 1,\dots,n$, $V_2^i(t) := \langle M^i(t)\rangle$ and, for $k \geq 3$, we define $V_k^i(t)$ as the compensator of the k-variation process $\sum_{s\leq t}|\Delta M^i(t)|^k$ of $M^i(t)$ (see Equation $(A3)$ on page 1795 in van de Geer (1995)).

In our case, we have, $n^{-1}\sum_{i=1}^n \int_0^\tau h(X_i,z)^2 dV_2^i(z) \leq \|h\|_n^2\|\alpha\|_{\infty,A}$, so that

$$S_\tau \leq \frac{n\delta^2\|\alpha\|_{\infty,A}}{\|h\|^2_{\infty,A}}\Big(\exp\big(a\|h\|_{\infty,A}\big) - 1 - a\|h\|_{\infty,A}\Big),$$

see the proof of Corollary 2.3. of van de Geer (1995). This majoration together with the proof of Lemma 2.2. in van de Geer (1995) yields the Bennett-type deviation inequality in our lemma. To obtain (42) and (43), we use the fact that $g(x) \geq 3x^2/(2(x+3))$ for any $x \geq 0$ and $g(x) \geq g_2(x)$ for any $x \geq 0$ where $g_2(x) := x+1-\sqrt{1+2x}$ and $g_2^{-1}(y) = \sqrt{2y}+y$, see Birgé and Massart (1998) p.366-367. □

The next Proposition is obtained from (43) by using a recent $L^2(\mu)-L^\infty$ generic chaining type of technique (see Talagrand (2005) and Baraud (2010)). This method is close to other $L^2(\mu)-L^\infty$ chaining methods, see among others Proposition 4 p. 282-287 in Comte (2001), Theorem 5 in Birgé and Massart (1998) and Proposition 7, Theorem 8 and Theorem 9 in Barron et al. (1999).

Proposition 4. *Let $\bar{S}$ be a D-dimensional linear subspace of $L^2 \cap L^\infty(\mu)$, and define B_δ as the $L^2(\mu)$ closed ball of $\bar{S}$ with radius δ. The L^∞-index of $\bar{S}$ is defined in the following way:*

$$\bar{r} = \frac{1}{\sqrt{D}} \inf_{(\psi_\lambda)} \sup_{\beta \neq 0} \frac{\| \sum_{\lambda \in \Lambda} \beta_\lambda \psi_\lambda \|_{\infty, A}}{|\beta|_\infty}, \tag{44}$$

where the infimum is taken over every orthonormal basis $(\psi_\lambda)_{\lambda \in \Lambda}$ of $\bar{S}$, and where $|\beta|_\infty$ is the ℓ_∞-norm of $\beta \in \mathbb{R}^\Lambda$. For any $x > 0$ and $\delta > 0$, we have

$$\mathbb{P}_{\Delta_\rho}\Big[\sup_{h \in B_\delta} \nu_n(h) \geq \kappa_0 \Big(\delta_\rho \sqrt{\frac{\|\alpha\|_{\infty, A}(D+x)}{n}} + \delta_\rho \bar{r} \frac{D+x}{n} \Big) \Big] \leq e^{-x},$$

where $\kappa_0 = 11.8$, and where we recall that $\delta_\rho = \delta(2 - 1/\rho)$, where $\rho > 1$ and where Δ_ρ is defined in (37).

Proof. Let us construct a sequence of partitions $(\mathcal{A}_k)_{k \geq 0}$ of B_δ with the following properties:

(1) It is increasing, in the sense that any element of $\mathcal{A}_{k+1}$ is included in an element of $\mathcal{A}_k$ [note that this is the key property of the generic chaining argument, see Talagrand (2005)].
(2) It is such that $\mathcal{A}_0 = \{B_\delta\}$ and for each $k \geq 1$, one has $|\mathcal{A}_k| \leq (2\pi e)^{D/2}(1 + 2^k)^D$.
(3) The diameters of the elements of $\mathcal{A}_k$ are controlled in the following way:

$$\operatorname{diam}(A, L^2(\mu)) \leq 2^{-k}\delta \quad \text{and} \quad \operatorname{diam}(A, L^\infty) \leq \bar{r} 2^{-k} \delta \quad \forall A \in \mathcal{A}_k \, \forall k \geq 1,$$

where we recall that $\operatorname{diam}(A, d) = \sup_{a,b \in A} d(a, b)$.

The construction of such a partition follows the construction given in Proposition 1 in Birgé and Massart (1998), or in Lemma 7.14 from Massart (2007). Below, we denote by $|x|_2$ and $|x|_\infty$ the ℓ_2 and ℓ_∞ norms of $x \in \mathbb{R}^D$. Without loss of generality we can assume that $\bar{r} = \bar{r}(\psi)$ for some orthonormal basis $\psi = (\psi_\lambda)_{\lambda \in \Lambda}$. Using the natural isometry between $\mathbb{R}^D$ and $\bar{S}$, we first construct a sequence $(\mathcal{A}'_k)_{k \geq 0}$ of partitions of $B'_\delta = \{\beta \in \mathbb{R}^D : |\beta|_2 \leq \delta\}$ with suitable properties. First, put $\mathcal{A}'_0 = \{B'_\delta\}$. Then, construct $\mathcal{A}'_1$ in the following way. Consider disjoint cubes $\{C'_1, \ldots, C'_N\}$ with vertices $\delta/\sqrt{D}$ that cover B'_δ, one of them being centered at 0. The partition $\mathcal{A}'_1$ is then simply given by the $C'_j \cap B'_\delta$. Then, in order to construct $\mathcal{A}'_{k+1}$ from $\mathcal{A}'_k$, simply partition each cube in $\mathcal{A}'_k$ using cubes with vertices half as small (and then equal to $2^{-(k+1)}\delta/\sqrt{D}$), and only keep the ones that have a nonempty intersection with B'_δ. By doing so, one has

$$\operatorname{diam}(C, \ell_\infty) \leq 2^{-k}\delta/\sqrt{D} \text{ and } \operatorname{diam}(C, \ell_2) \leq 2^{-k}\delta \quad \forall C \in \mathcal{A}'_k,$$

and this construction entails that the sequence of partition $(\mathcal{A}'_k)_{k \geq 0}$ is increasing. Moreover, the volumetric argument from the proof of Lemma 7.14 in Massart (2007), gives

$$|\mathcal{A}'_k| \leq (2\pi e)^{D/2}(1 + 2^k)^D \quad \forall k \geq 0.$$

It is now easy to construct $(\mathcal{A}_k)_{k \geq 0}$. For each $k \geq 1$, if $\mathcal{A}'_k := \{C'_1, \ldots, C'_N\}$, consider

$$C_j := \Big\{ \sum_{\lambda \in \Lambda} \beta_\lambda \psi_\lambda : (\beta_\lambda) \in C'_j \Big\},$$

and simply put

$$\mathcal{A}_k := \Big\{ C_j - \Big(\bigcup_{1 \le i < j} C_i \Big) \cap B_\delta : 1 \le j \le N \Big\},$$

with the convention that $\cup_\emptyset = \emptyset$. This provides a sequence of partitions with the properties (1)-(3), thanks to the definition of $\bar{r}$[1].

Now, we want to use the generic chaining type of argument (see Talagrand (2005)), that was proposed in Baraud (2010), see the proofs of Theorems 2.1 and 5.1 therein. The difference between what we do here and the proof of Baraud (2010) is minor: here the L^2 and L^∞ norms are explicit, so we are able to take advantage of the covering by cubes (see above) while Theorems 2.1 and 5.1 in Baraud (2010) are move general, since they hold for any distances (so two partitions and two volumetric arguments are used, while this can be avoided here).

For any $k \ge 1$ and $A \in \mathcal{A}_k$, fix an arbitrary element $h_k(A)$. Then, for any $h \in B_\delta$, define $\pi_k(h)$ in the following way: take the unique $A \in \mathcal{A}_k$ such that $h \in A$, and define $\pi_k(h) = h_k(A)$. Define also $\pi_0(h) = 0$ (since $0 \in B_\delta$). Now, for any $h \in B_\delta$, the following decomposition holds:

$$\nu_n(h) = \sum_{k \ge 0} \Big(\nu_n(\pi_{k+1}(h)) - \nu_n(\pi_k(h)) \Big),$$

so, if $z = \sum_{k \ge 0} z_k$ where z_k are positive numbers, we have

$$\mathbb{P}_\Delta \Big(\sup_{h \in B_\delta} \nu_n(h) \ge z \Big) \le \sum_{k \ge 0} \sum_{(s,u) \in E_k} \mathbb{P}_\Delta \Big(\nu_n(s) - \nu_n(u) \ge z_k \Big),$$

where $E_k := \{ (\pi_k(h), \pi_{k+1}(h)) : h \in B_\delta \}$. It must be noted that, since the partitions $(\mathcal{A}_k)_{k \ge 0}$ are increasing, both $\pi_k(h)$ and $\pi_{k+1}(h)$ belong to the same element of $\mathcal{A}_k$ for any $h \in B_\delta$, so $\|s-u\|_\infty \le \bar{r}\delta_\rho 2^{-k}$ and $\|s-u\|_\mu \le \delta_\rho 2^{-k}$ for any $(s,u) \in E_k$. Moreover, $\pi_{k+1}(h)$ uniquely determines $\pi_k(h)$, so that $|E_k| \le |\mathcal{A}_{k+1}| = N_k$ where $N_k := (2\pi e)^{D/2}(1+2^{k+1})^D$. Consider

$$z_k := 2^{-k}\delta_\rho \sqrt{\frac{2\|\alpha\|_\infty x_k}{n}} + 2^{-k}\bar{r}\delta_\rho \frac{x_k}{n} \text{ where } x_k = x + \log(2^{k+1}N_k).$$

Using (43), one has for any $(s,u) \in E_k$:

$$\mathbb{P}_\Delta \Big(\nu_n(s) - \nu_n(u) \ge z_k \Big) \le (2^{k+1}N_k)^{-1} e^{-x},$$

so

$$\mathbb{P}_\Delta \Big(\sup_{h \in B_\delta} \nu_n(h) \ge z \Big) \le e^{-x},$$

and an easy computation shows that

$$z = \sum_{k \ge 0} z_k \le \kappa_0 \Big(\delta_\rho \sqrt{\frac{\|\alpha\|_\infty (D+x)}{n}} + \delta_\rho \bar{r} \frac{D+x}{n} \Big),$$

where $\kappa_0 = 11.8$. This concludes the proof of Proposition 4. □

[1]The only minor difference between this construction and the constructions of Birgé and Massart (1998) and Massart (2007) is that here the center of the cubes are such that the embedding $\mathcal{A}_{k+1} \subset \mathcal{A}_k$ holds.

Now, we can turn to the proof of Proposition 3. We denote by $D(m,m')$ the dimension of the linear space $S_m + S'_m$.

Proof of Proposition 3. In Proposition 4, take $x = D_{m'} + u$, $\delta = 1$, $B_\delta = B^\mu_{m,m'}(0,1) = \{t \in S_m + S_{m'} : \|t\|_\mu \leq 1\}$ and $\rho = 2$ in order to get:

$$\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m,m'}(0,1)} \nu_n^2(h) \geq \eta^2\Big] \leq 2\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m,m'}(0,1)} \nu_n(h) \geq \eta\Big] \leq 2e^{-D_{m'}-u},$$

where

$$\begin{aligned}
\eta^2 &= \frac{9}{4}\kappa_0^2\Big(\sqrt{\|\alpha\|_{\infty,A}\frac{D(m,m')+D_{m'}+u}{n}} + \bar r_{m,m'}\frac{D(m,m')+D_{m'}+u}{n}\Big)^2\\
&\leq \frac{9}{2}\kappa_0^2\Big(\|\alpha\|_{\infty,A}\frac{D(m,m')+D_{m'}+u}{n} + 2\bar r^2_{m,m'}\big(\frac{D(m,m')+D_{m'}}{n}\big)^2 + 2\bar r^2_{m,m'}\frac{u^2}{n^2}\Big)\\
&\leq 18\kappa_0^2\Big((1+\|\alpha\|_{\infty,A})\frac{D_m+D_{m'}}{n} + \big(\frac{\|\alpha\|_{\infty,A}u}{n} \vee \bar r^2_{m,m'}\frac{u^2}{n^2}\big)\Big),
\end{aligned}$$

where we used the fact that

$$\bar r^2_{m,m'}\big(\frac{D(m,m')+D_{m'}}{n}\big)^2 \leq \frac{D(m,m')}{n},$$

for n large enough (see Appendix B) and $D(m,m') \leq D_m + D_{m'}$. This gives

$$\begin{aligned}
&\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m,m'}(0,1)} \nu_n^2(h) \geq \kappa\Big((1+\|\alpha\|_{\infty,A})\frac{D_m+D_{m'}}{n} + \big(\frac{\|\alpha\|_{\infty,A}u}{n} \vee \bar r^2_{m,m'}\frac{u^2}{n^2}\big)\Big)\Big]\\
&\qquad\qquad \leq 2e^{-D_{m'}-u}, \qquad (45)
\end{aligned}$$

where $\kappa = 18\kappa_0^2$. Now, we set $p(m,m') = \kappa(1+\|\alpha\|_{\infty,A})(D_m + D_{m'})/n$ with $\kappa = 18\kappa_0^2$. This gives

$$\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m,m'}(0,1)} \nu_n^2(h) \geq p(m,m') + z\Big] \leq \begin{cases} 2e^{-D_{m'}-nz/(\kappa\|\alpha\|_{\infty,A})} & \text{if } z \leq \kappa\|\alpha\|^2_{\infty,A}/\bar r^2_{m,m'}\\ 2e^{-D_{m'}-n\sqrt{z}/\sqrt{\kappa\bar r^2_{m,m'}}} & \text{if } z > \kappa\|\alpha\|^2_{\infty,A}/\bar r^2_{m,m'}, \end{cases}$$

and we obtain that

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sup_{h\in B^\mu_{m,m'}(0,1)} \nu_n^2(h) - p(m,m')\Big)_+ \mathbb{1}(\Delta)\Big]\\
&\qquad\leq \int_0^\infty \mathbb{P}_\Delta\Big(\sup_{h\in B^\mu_{m,m'}(0,1)} \nu_n^2(h) > p(m,m') + z\Big)dz\\
&\qquad\leq 2e^{-D_{m'}}\Big(\int_0^\infty e^{-nz/(\kappa\|\alpha\|_{\infty,A})}dz + \int_0^{+\infty} e^{-n\sqrt{z}/\sqrt{\kappa\bar r^2_{m,m'}}}dz\Big)\\
&\qquad\leq 2e^{-D_{m'}}\frac{\kappa}{n}\Big(\|\alpha\|_{\infty,A}\int_0^\infty e^{-v}dv + \frac{\bar r^2_{m,m'}}{n}\int_0^\infty e^{-\sqrt{v}}dv\Big)\\
&\qquad\leq 2e^{-D_{m'}}\frac{\kappa}{n}(\|\alpha\|_{\infty,A} + \frac{2\bar r^2_{m,m'}}{n}) \leq \frac{\kappa_\alpha e^{-D_{m'}}}{n},
\end{aligned}$$

where we used the upper bounds of $\bar{r}_{m,m'}$ given in Appendix B and where κ_α is a constant depending on $\|\alpha\|_{\infty,A}, f_0$ and the basis. It remains to bound from above $\sum_{m'\in\mathcal{M}_n} e^{-D_{m'}}$. This term is at most

$$\sum_{j,k\geq 1} e^{-jk} = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}(e^{-j})^k = \sum_{j=1}^{\infty}\frac{e^{-j}}{1-e^{-j}} \leq \frac{1}{1-e^{-1}}\sum_{j=1}^{\infty}e^{-j} = \frac{e^{-1}}{(1-e^{-1})^2}.$$

This concludes the proof of Proposition 3 when n is large enough. The statement of Proposition 3 is obvious for small n, up to an increased constant C_2. □

6.5. **Proof of Lemma 2.** First, we write

$$\begin{aligned}\mathbb{E}[\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha\|_A^2 \mathbb{1}(\Lambda\cap\Delta\cap\Omega)^{\complement})] &\leq \mathbb{E}[\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha\|_A^2 \mathbb{1}(\Delta\cap\Omega)^{\complement})] \\ &\quad + \mathbb{E}[\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha\|_A^2 \mathbb{1}(\Lambda^{\complement}\cap(\Delta\cap\Omega))].\end{aligned}$$

The first term is bounded by C_1/n as in the proof of Theorem 1 by using Lemma 3, Lemma 4 and Assumption $(\mathcal{A}5)$. For the second term, we get

$$\mathbb{E}[\|\hat{\alpha}_{\hat{\hat{m}}} - \alpha\|_A^2 \mathbb{1}(\Lambda^{\complement}\cap(\Delta\cap\Omega))] \leq 8(\mathbb{E}^{1/2}(\|\hat{\alpha}_{\hat{\hat{m}}}\|^4) + \|\alpha\|_A^2)\big(\mathbb{P}(\Lambda^{\complement}\cap(\Delta\cap\Omega))\big)^{1/2}.$$

Lemma 3 can be applied again: this gives $\mathbb{E}(\|\hat{\alpha}_{\hat{\hat{m}}}\|^4) \leq C_B n^5$, so we have to prove that

$$\mathbb{P}(\Lambda^{\complement}\cap(\Delta\cap\Omega)) \leq C_k n^{-k}, \tag{46}$$

for $k \geq 7/2$. Let us do the decomposition

$$\begin{aligned}\mathbb{P}(\Lambda^{\complement}\cap(\Delta\cap\Omega)) &= \mathbb{P}\left[(|\|\hat{\alpha}_{m^*}\|_\infty - \|\alpha\|_{\infty,A}| > \|\alpha\|_{\infty,A}|)\cap(\Delta\cap\Omega)\right] \\ &\leq \mathbb{P}\left[(\|\hat{\alpha}_{m^*} - \alpha\|_{\infty,A} > \|\alpha\|_{\infty,A}/2)\cap(\Delta\cap\Omega)\right] \\ &\leq \mathbb{P}\left[(\|\hat{\alpha}_{m^*} - \alpha_{m^*}\|_{\infty,A} > \|\alpha\|_{\infty,A}/4)\cap(\Delta\cap\Omega)\right] \\ &\quad + \mathbb{P}\left[(\|\alpha_{m^*} - \alpha\|_{\infty,A} > \|\alpha\|_{\infty,A}/4)\cap(\Delta\cap\Omega)\right].\end{aligned}$$

Assumption $(\mathcal{M}2)$ implies that $\|\hat{\alpha}_{m^*} - \alpha_{m^*}\|_{\infty,A} \leq \phi_0\sqrt{D_{m_1^*}D_{m_2^*}}\|\hat{\alpha}_{m^*} - \alpha_{m^*}\|_A$. Moreover, it follows from (23) in the proof of Theorem 1, that we have, on $\Delta\cap\Omega$:

$$\|\hat{\alpha}_{m^*} - \alpha_{m^*}\|_A^2 \leq 32\Big(\|\alpha\mathbb{1}(A) - \alpha_{m^*}\|_n^2 + \sup_{h\in B_{m^*}^{\mu}(0,1)}\nu_n^2(h)\Big).$$

As $\|h\|_n^2 \leq \tau\|h\|_{\infty,A}^2$ for any h supported by A, we get

$$\begin{aligned}\mathbb{P}(\Lambda^{\complement}\cap(\Delta\cap\Omega)) &\leq \mathbb{P}\Big[(32\phi_0^2 D_{m_1^*}D_{m_2^*}\sup_{h\in B_{m^*}^{\mu}(0,1)}\nu_n^2(h) > \|\alpha\|_{\infty,A}^2/32)\cap(\Delta\cap\Omega)\Big] \\ &\quad + \mathbb{P}\Big[\phi_0\sqrt{D_{m_1^*}D_{m_2^*}}\|\alpha_{m^*} - \alpha\|_{\infty,A} > \|\alpha\|_{\infty,A}/(32\sqrt{\tau})\Big] \\ &\quad + \mathbb{P}\Big[\|\alpha_{m^*} - \alpha\|_{\infty,A} > \|\alpha\|_{\infty,A}/4\Big].\end{aligned}$$

The last two probabilites are studied with the same Nikol'skii (1975)'s argument as in the proof of Proposition 1, for n large enough and using the assumption that $\beta_1 > 2, \beta_2 > 2$. Since $D_{m_1^*}D_{m_2^*} \leq \sqrt{n}$, we can bound the first probability from above by

$$\mathbb{P}\Big[\Big\{\sup_{h\in B_{m^*}^{\mu}(0,1)}\nu_n^2(h) > \frac{\|\alpha\|_{\infty,A}^2}{2^{10}\phi_0^2\sqrt{n}}\Big\}\cap\Delta\cap\Omega\Big]. \tag{47}$$

Using (45) with the upper bound of $\bar{r}_{m,m'}$ for the collection [T] which is given in Appendix B, we obtain that for $\kappa_\alpha := \kappa(1 + \|\alpha\|_{\infty,A})$:

$$\mathbb{P}_\Delta\Big[\sup_{h\in B^\mu_{m^*}(0,1)} \nu_n^2(h) > \kappa_\alpha\Big(\frac{D_{m^*}}{n} + (\frac{u}{n} + \frac{\sqrt{n}u^2}{n^2})\Big)\Big] \le 2e^{-u}.$$

So, taking $u = \|\alpha\|_{\infty,A}\sqrt{n}/(2^{10}\phi_0^2)$ and since $D_{m^*} \le \sqrt{n}$, we obtain that (47) is smaller than $2e^{-\|\alpha\|_{\infty,A}\sqrt{n}/(2^{10}\phi_0^2)}$. This ensures (46) for any integer k and concludes the proof of Lemma 2. □

Appendix A. Some useful tools from wavelet and approximation theory

A.1. **The basis** $[W]$**.** Consider a pair $\{\phi, \psi\}$ of scaling function and wavelet, where ψ has K vanishing moments. Then ϕ and ψ have a support width of at least $2K-1$, and there is a pair with minimal support, see Daubechies (1988). This is the starting point of the construction of an orthonormal wavelet basis of $\mathbb{L}^2[0,1]$, as proposed in Cohen et al. (1993). Roughly, the idea is to retain the interior scaling functions (those "far" from the edges 0 and 1), and to add adapted edge scaling functions. This is done in Cohen et al. (1993), see Section 4 and Theorem 4.4, where the construction allows to keep the orthonormality of the system and the number of vanishing moment unchanged, as well as the number 2^j of scaling function at each resolution j (which improves a previous construction by Meyer (1991)). Indeed, if l is such that $2^l \ge 2K$, consider for $j \ge l-1$:

$$\Psi_{j,k} := \begin{cases} \psi^0_{j,k} & \text{if } j \ge l \text{ and } k = 0,\dots,K-1 \\ \psi_{j,k} & \text{if } j \ge l \text{ and } k = K,\dots,2^j-K-1 \\ \psi^1_{j,k} & \text{if } j \ge l \text{ and } k = 2^j-K,\dots,2^j-1 \\ \phi^0_{l,k} & \text{if } j = l-1 \text{ and } k = 0,\dots,K-1 \\ \phi_{l,k} & \text{if } j = l-1 \text{ and } k = K,\dots,2^l-K-1 \\ \phi^1_{l,k} & \text{if } j = l-1 \text{ and } k = 2^l-K,\dots,2^l-1 \end{cases}$$

where $\phi_{j,k} = 2^{j/2}\phi(2^j\cdot - x)$ and $\psi_{j,k} = 2^{j/2}\psi(2^j\cdot - x)$ are the "interior" dilatations and translations of $\{\phi,\psi\}$, and $\phi^0_{j,k}, \psi^0_{j,k}, \phi^1_{j,k}, \psi^1_{j,k}$ are, at each resolution j, dilatations of $2K$ edge scaling functions and wavelets (K for each edge). We know from Cohen et al. (1993) that the collection

$$W := \{\Psi_{j,k} : j \ge l-1, k = 0,\dots,2^j-1\}$$

is an orthonormal basis of $\mathbb{L}^2[0,1]$, and the interior and edge wavelets have K vanishing moments, which ensures that the elements of this collection have the same smoothness as ϕ and ψ.

A.2. **Some approximation results.** An orthonormal basis of $\mathbb{L}^2[0,1]^2$ is simply obtained by taking tensor products of two bases $[W]$ for instance. If $W^{(1)}$ and $W^{(2)}$ are two basis $[W]$ (we can use two different pairs $\{\phi^{(1)},\psi^{(1)}\}$ and $\{\phi^{(2)},\psi^{(2)}\}$ with possibly different number of vanishing moments), we can simply consider

$$W^{(1)} \otimes W^{(2)} := \{\Psi^{(1)}_{j_1,k_1} \otimes \Psi^{(2)}_{j_2,k_2} : j_1 \ge l_1 - 1, j_2 \ge l_2 - 1, \\ k_1 = 0,\dots,2^{j_1}-1, k_2 = 0,\dots,2^{j_2}-1\},$$

where $\Psi^{(1)}_{j_1,k_1} \otimes \Psi^{(2)}_{j_2,k_2}(x_1,x_2) := \Psi^{(1)}_{j_1,k_1}(x_1)\Psi^{(2)}_{j_2,k_2}(x_2)$. We can also obtain an orthonormal basis of $\mathbb{L}^2[0,1]^2$ by taking tensor products of two collections among the ones considered in Section 2.4. Let us consider S_m as one of the following:

- A space of piecewise polynomials (see Section 2.4, basis $[DP]$) of degrees smaller than $s_i > \beta_i - 1$ $(i = 1, 2)$ based on a partition with rectangles of sidelengthes $1/D_{m_1}$ and $1/D_{m_2}$;
- A space spanned by tensors products of $[W]$, namely the span of the $\Psi^{(1)}_{j_1,k_1} \otimes \Psi^{(2)}_{j_2,k_2}$ for $j_1 \in \{l-1,\dots,m_1\}$, $j_2 \in \{l-1,\dots,m_2\}$, $k_1 \in \{0,\dots,2^{j_1}-1\}$, $k_2 \in \{0,\dots,2^{j_2}-1\}$, where the $\Psi^{(1)}_{j,k}$ and $\Psi^{(2)}_{j,k}$ have respective regularities $s_1 > \beta_1 - 1$ and $s_2 > \beta_2 - 1$ (here $D_{m_i} = 2^{m_i}, i = 1, 2$);
- The space of trigonometric polynomials with degree smaller than D_{m_1} in the first direction and smaller than D_{m_2} in the second direction.

Note that the dimension of each space is $D_{m_1}D_{m_2}$. The following result is an easy consequence of results by Hochmuth (2002) and Nikol'skii (1975) (see Lacour (2007a)).

Lemma 6. *Let s belong to $B^{\boldsymbol{\beta}}_{2,\infty}(A)$ where $\boldsymbol{\beta} = (\beta_1, \beta_2)$. We consider that S_m is one of the spaces above, with dimension $D_{m_1}D_{m_2}$. If s_m is the orthogonal projection of s on S_m, then there is a positive constant C such that*

$$\|s - s_m\|_A = \Big(\int_A |s - s_m|^2 \Big)^{1/2} \leq C[D_{m_1}^{-\beta_1} + D_{m_2}^{-\beta_2}],$$

where C depends on the Besov norm of s and on the basis.

Appendix B. Upper bounds for the L^∞-index

In this section we provide controls on the L^∞-index $\bar{r} = \bar{r}_{m,m'}$ of $\bar{S} = S_m + S_{m'}$ (which is defined in Proposition 4, see Section 6.4). Recall that each S_m is a tensor product model, which can be spanned by any of the basis $[DP]$, $[T]$ or $[W]$, see Section 2.4. Recall that $B^{\mu}_{m,m'}(0,1) = \{t \in S_m + S_{m'}, \|t\|_\mu \leq 1\}$, that $S_m + S_{m'} \subset \mathcal{S}_n$ and that the norm connection holds, see Condition $(\mathcal{M}_2)$ in Section 2.4. We denote for short $D(m,m') = \dim(S_m + S_{m'})$. We give below upper bounds for $\bar{r}_{m,m'}$ and for

$$\bar{r}^2_{m,m'} \Big(\frac{D(m,m')}{n} + \frac{D_{m'}}{n} \Big)^2,$$

which is a quantity that appears in the proof of Proposition 3.

B.1. Collection [DP].

As $S_m + S_{m'}$ is a linear space, an orthonormal $L^2(\mu)$-basis $(\psi_\lambda)_{\lambda \in \Lambda_n}$ can be built by orthonormalisation on each sub-rectangle of $(\varphi_\lambda)_{\lambda \in \Lambda_n}$, the orthonormal basis of $\mathcal{S}_n$. We denote by r_1 (respectively r_2) the maximal degree in the x-direction (resp.

in the y-direction). Then

$$\begin{aligned}\sup_{\beta\neq 0}\frac{\|\sum_{\lambda\in\Lambda_n}\beta_\lambda\psi_\lambda\|_{\infty,A}}{|\beta|_\infty} &\leq \|\sum_{\lambda\in\Lambda_n}|\psi_\lambda|\|_{\infty,A}\leq (r_1+1)(r_2+1)\sup_{\lambda\in\Lambda_n}\|\psi_\lambda\|_{\infty,A}\\ &\leq \phi_0(r_1+1)(r_2+1)\sqrt{N_n}\sup_{\lambda\in\Lambda_n}\|\psi_\lambda\|\\ &\leq \phi_0(r_1+1)(r_2+1)\sqrt{N_n}\sup_{\lambda\in\Lambda_n}\|\psi_\lambda\|_\mu/\sqrt{f_0}\\ &\leq \phi_0(r_1+1)(r_2+1)\sqrt{N_n/f_0},\end{aligned}$$

thus

$$\bar{r}_{m,m'}\leq\frac{\phi_0(r_1+1)(r_2+1)\sqrt{N_n}}{\sqrt{f_0D(m,m')}}.$$

Moreover, since $N_n\leq n/\log n$ and $D_{m'}\leq D(m,m')$, we have

$$\begin{aligned}\bar{r}^2_{m,m'}\Big(\frac{D(m,m')}{n}+\frac{D_{m'}}{n}\Big)^2 &\leq \frac{2\phi_0^2(r_1+1)^2(r_2+1)^2N_n}{f_0D(m,m')}\frac{D(m,m')^2}{n^2}\\ &\leq \frac{2\phi_0^2(r_1+1)^2(r_2+1)^2}{f_0\log n}\frac{D(m,m')}{n}\leq\frac{D(m,m')}{n}\end{aligned}$$

for n large enough.

B.2. **Collection [T].** For trigonometric polynomials, we write for $\beta\neq 0$:

$$\frac{\|\sum_{\lambda\in\Lambda_n}\beta_\lambda\psi_\lambda\|_{\infty,A}}{|\beta|_\infty}\leq\frac{\phi_0\sqrt{N_n}\|\sum_\lambda\beta_\lambda\psi_\lambda\|_\mu}{\sqrt{f_0}|\beta|_\infty}\leq\frac{\phi_0\sqrt{N_n}\sqrt{\sum_\lambda\beta_\lambda^2}}{\sqrt{f_0}|\beta|_\infty}\leq\frac{\phi_0\sqrt{N_nD(m,m')}}{\sqrt{f_0}},$$

so that $\bar{r}_{m,m'}\leq\phi_0\sqrt{N_n/f_0}$. Moroever, since $N_n\leq\sqrt{n}/\log n$, we obtain

$$\bar{r}^2_{m,m'}\Big(\frac{D(m,m')}{n}+\frac{D_{m'}}{n}\Big)^2\leq\frac{2\phi_0^2N_nD(m,m')^2}{f_0n^2}\leq\frac{2\phi_0^2}{f_0\log n}\frac{D(m,m')}{n}\leq\frac{D(m,m')}{n}$$

for n large enough.

B.3. **Collection [W].** Recall that this collection is precisely defined in Appendix A. It suffices to use the fact that for $j_1\geq l_1$ and $j_2\geq l_2$ fixed, the tensor products $\Psi^{(1)}_{j_1,k_1}\otimes\Psi^{(2)}_{j_2,k_2}$ have disjoint supports, expected for a finite number of indexes k_1, k_2, that depends only on the support of the scaling and mother wavelet functions used in the construction of the basis $[W]$, for both dimensions. Then, using as for the case [DP] the embedding $S_m+S_{m'}\subset\mathcal{S}_n$, and if $\{\Psi_\lambda:\lambda\in\Lambda_n\}$ is the collection that spans $\mathcal{S}_n$ (the one with the largest dimension in each direction), we obtain

$$\|\sum_{\lambda\in\Lambda_n}\beta_\lambda\Psi_\lambda\|_{\infty,A}\leq C(\Psi^{(1)},\Psi^{(2)})\sqrt{N_n}|\beta|_\infty,$$

where $C(\Psi^{(1)},\Psi^{(2)})$ is a constant that depends only the scaling and mother wavelet functions used in the construction of the basis, and not on the resolution level. Hence,

$\bar{r}_{m,m'} \leq C(\Psi^{(1)}, \Psi^{(2)})\sqrt{N_n/D(m,m')}$. Moreover, since $N_n \leq n/\log n$ we obtain:

$$\bar{r}^2_{m,m'}\Big(\frac{D(m,m')}{n} + \frac{D_{m'}}{n}\Big)^2 \leq \frac{2C(\Psi^{(1)}, \Psi^{(2)})^2 N_n}{D(m,m')}\frac{D(m,m')^2}{n^2}$$
$$\leq 2\frac{C(\Psi^{(1)}, \Psi^{(2)})^2}{\log(n)}\frac{D(m,m')}{n} \leq \frac{D(m,m')}{n},$$

for n large enough.

References

P. K. Andersen, O., Borgan, R. D. Gill and N. Keiding (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York.

Y. Baraud (2010) A Bernstein-type inequality for suprema of random processes with an application to statistics. To appear in *Bernoulli*.

Y. Baraud and L. Birgé (2009). Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Relat. Fields*, 149(1-2), 239–284.

Y. Baraud, F. Comte and G. Viennet (2001). Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.*, 29(3):839-875.

Y. Baraud, F. Comte and G. Viennet (2001). Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.* 5, 33-49.

A. Barron, L. Birgé and P. Massart (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113(3):301-413.

J. Beran (1981). Nonparametric regression with randomly censored survival data. Technical report, Dept. Statist. Univ. California, Berkeley.

L. Birgé and P. Massart (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* 4(3):329-375.

E. Brunel, F. Comte and C. Lacour (2007). Adaptive estimation of the conditional density in presence of censoring. *Sankhya* 69(4):734-763.

G. Castellan and F. Letué (2000). Estimation of the Cox regression function via model selection. *Chapter of the PhD thesis of F. Letué*, University Paris XI-Orsay.

A. Cohen, I. Daubechies and P.B. Vial (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* 1(1):54-81.

F. Comte (2001). Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli*, 7(2):267-298.

D.R. Cox (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B*, 34, 187-220.

D.M. Dabrowska (1987). Nonparametric regression with censored survival time data. *Scand. J. Statist.*, 14(3):181-197.

D.M. Dabrowska (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. Ann. Statist. 17(3):1157-1167.

I. Daubechies (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* 41(7):909-996.

M. Delecroix, O. Lopez and V. Patilea (2008). Nonlinear censored regression using synthetic data. *Scand. J. Statist.* 35 (2):248-265.

G. Grégoire (1993). Least squares cross-validation for counting processes intensities. *Scand. J. Statist.*, 20(4):343-360.

C. Heuchenne and I. Van Keilegom (2007). Location estimation in nonparametric regression with censored data. *J. Multivariate Anal.*, 98(8):1558-1582.

R. Hochmuth (2002). Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179-208.

J. Huang (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.*, 27(5):1536-1563.

M. Jacobsen (1982). *Statistical analysis of counting processes*. Lecture Note in Statistics 12. Springer-Verlag, New York.

A.F. Karr (1986). *Point processes and their statistical inference*. Probability: Pure and Applied. Marcel Dekker Inc. New York.

W. Härdle, G. Kerkyacharian, D. Picard and A. Tsybakov (1998). *Wavelets, approximation, and statistical applications.* Lecture Notes in Statistics, 129. Springer-Verlag, New York.

C. Lacour (2007a). Adaptive estimation of the transition density of a markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(5):571-597.

C. Lacour (2007b). Estimation non paramétrique adaptative pour les chaînes de Markov et les chaînes de Markov cachées. *PhD thesis*. http://www.math.u-psud.fr/ lacour/etudes/

M. LeBlanc and J. Crowley (1999). Adaptive regression splines in the Cox model. *Biometrics*, 55, 204-213.

G. Li and H. Doss (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.*, 23(3):787-823.

O. B. Linton, J. P. Nielsen and S. Van de Geer (2003). Estimating the multiplicative and additive hazard fonctions by kernel methods. *Ann. Statist.*, 31(2):464-492.

R. S. Liptser and A. N. Shiryayev (1989). *Theory of martingales*, vol. 49 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht. Translated from the Russian by K. Dzjaparidze [Kacha Dzhaparidze].

P. Massart (2007). *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

I. W. McKeague and K. J. Utikal (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.*, 18(3):1172-1187.

Y. Meyer (1991). Ondelettes sur l'intervalle. *Revista Matemática Iberoamericana* 7(2):115-133.

S. M. Nikol'skii (1975). *Approximation of functions of several variables and imbedding theorems.* Springer-Verlag, New York. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

H. Ramlau-Hansen (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, 11(2):453-466.

P. Reynaud-Bouret (2003). Adaptive estimation of the intensity of nonhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Relat. Fields*, 126(1):103-153.

P. Reynaud-Bouret (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4): 633-661.

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6), 1348-1360.

W. Stute (1986). Conditional empirical processes. *Ann. Statist.*, 14(2): 638-647.

W. Stute (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, 23(4): 461-471.

M. Talagrand (1996). New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505-563.

Talagrand, M. (2005). *The generic chaining*. Springer Monographs in Mathematics, Springer-Verlag, Berlin. Upper and lower bounds of stochastic processes.

H. Triebel (2006). *Theory of function spaces. III.* Monographs in Mathematics, 100. Birkhäuser Verlag, Basel, 2006.

A. Tsybakov (2003a). *Introduction à l'estimation non-paramétrique*. Springer.

S. van de Geer (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, 23(5):1779–1801.